School of Electrical and Computer Engineering

Implementation of a Platform for the Update, Management and Analysis of Data for the «HelTh» Nutrition Database

Diploma Thesis
in Electrical and Computer Engineering

by

Evangelos Stylianos Vlassopoulos

Committee Members:

Prof. Michail Zervakis (School of Electrical and Computer Engineering, Technical University of Crete)

Prof. Michail Lagoudakis (School of Electrical and Computer Engineering, Technical University of Crete)

Prof. Maria Kapsokefalou (Department of Food Science and Human Nutrition, Agricultural University of Athens)

Chania, Greece

June 2025

Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Η/Υ

Υλοποίηση Πλατφόρμας για την Ενημέρωση, Διαχείριση και Ανάλυση της Βάσης Διατροφικών Δεδομένων «HelTH»

Διπλωματική εργασία

στη **Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Η/Υ**

από τον

Ευάγγελο Στυλιανό Βλασσόπουλο

Committee Members:

Καθ. Μιχαήλ Ζερβάκης (Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Η/Υ, Πολυτεχνείο Κρήτης)

Καθ. Μιχαήλ Λαγουδάκης (Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Η/Υ, Πολυτεχνείο Κρήτης)

Καθ. Μαρία Καψοκεφάλου (Τμήμα Επιστήμης Τροφίμων & Διατροφής του Ανθρώπου, Γεωπονικό Πανεπιστήμιο Αθηνών)

Χανιά, Ελλάδα

Ιούνιος 2025

Abstract

Purpose: The study aims to test whether Natural Language Processing (NLP) and Machine Learning teaching can be employed to accurately predict the nutritional composition- namely total fat, protein, total sugar, sodium and fiber content- of food products using their ingredient list as input. This approach is centered around the development of AI-tool to support food labelling standardization, address public health concerns and raise consumer awareness.

Methodology: DistilBERT embeddings were employed to transform text from a food's ingredient list into structured numerical representation, in a deep learning based predictive framework. The experimental dataset was the USDA FoodData Central Branded Food Composition database which ensures a comprehensive representation of the food environment and the variation in composition. Experimental regression models and Multi-Layer Perceptron (MLP) networks employed a variety of loss functions, epochs, dataset sizes and batch sizes. The evaluation of the different experimental conditions was carried out using validation loss, Mean Absolute Error (MAE), and R2 score. Optimization was carried out using AdamW.

Results: Findings indicate that using datasets with data from a single food category (category-specific), provide models with improved predictive accuracy, validation loss and model convergence compared to those using data from various food categories (generalized). SmoothL1Loss function was associated with improved validation and training loss compared to other loss functions, while AdamW enhanced training stability. The study further highlights that using datasets with higher structure as opposed to unstructured datasets improves prediction accuracy and reduces noise and overfitting risks.

Conclusions: The results indicate that NLP-driven models can be proposed as a reliable alternative in the estimation/prediction of a food's nutritional composition from its ingredient list. This proposes the choice of scalable and cost-effective AI-based alternatives to traditional laboratory-based methods. Future research needs are identified in the areas of refinement of real-time prediction capabilities, optimization of feature selection techniques and ultimately the usability of such techniques in regulatory environments. The study highlights the potential of

machine learning and intelligent food composition prediction in the food industry as a tool to increase consumer trust and support high quality labelling.

Περίληψη

Σκοπός: Η μελέτη διερευνά τη δυνατότητα χρήσης της Επεξεργασίας Φυσικής Γλώσσας (NLP) και τεχνικών μηχανικής μάθησης για την ακριβή πρόβλεψη της διατροφικής σύστασης τροφίμων— ειδικότερα της περιεκτικότητας σε ολικά λιπαρά, πρωτεΐνες, ολικά σάκχαρα, νάτριο και φυτικές ίνες —βάσει της λίστας συστατικών τους. Η έρευνα υποκινείται από την αυξανόμενη ζήτηση για ακριβή και τυποποιημένη επισήμανση τροφίμων λόγω των ρυθμιστικών αλλαγών, των ανησυχιών για τη δημόσια υγεία και της αυξημένης καταναλωτικής ευαισθητοποίησης.

Μεθοδολογία: Αναπτύχθηκε ένα προγνωστικό πλαίσιο βασισμένο στη βαθιά μάθηση, αξιοποιώντας DistilBERT embeddings για τη μετατροπή των λιστών συστατικών σε αριθμητικές αναπαραστάσεις. Χρησιμοποιήθηκε ένα σύνολο δεδομένων από την USDA FoodData Central, διασφαλίζοντας την ευρεία κάλυψη της διατροφικής σύστασης τροφίμων. Πειραματικά μοντέλα παλινδρόμησης και δίκτυα Multi-Layer Perceptron (MLP), μελετησαν μια ποικιλία συναρτήσεων απώλειας, εποχών, μεγεθών συνόλου δεδομένων και μεγεθών παρτίδας. Η αξιολόγηση των διαφορετικών πειραματικών συνθηκών πραγματοποιήθηκε με τη χρήση της απώλειας επικύρωσης (validation loss), του μέσου απόλυτου σφάλματος (MAE) και του Συντελεστή Προσδιορισμού (R^2 Score). Η βελτιστοποίηση πραγματοποιήθηκε με τη χρήση του AdamW.

Αποτελέσματα: Τα ευρήματα δείχνουν ότι η χρήση συνόλων δεδομένων με δεδομένα από μία μόνο κατηγορία τροφίμων (ειδική κατηγορία), παρέχει μοντέλα με βελτιωμένη ακρίβεια πρόβλεψης, απώλεια επικύρωσης και σύγκλιση του μοντέλου σε σύγκριση με εκείνα που χρησιμοποιούν δεδομένα από διάφορες κατηγορίες τροφίμων (γενικευμένα). Η συνάρτηση SmoothL1Loss συσχετίστηκε με βελτιωμένες απώλειες επικύρωσης και εκπαίδευσης σε σύγκριση με άλλες συναρτήσεις απωλειών, ενώ η AdamW ενίσχυσε τη σταθερότητα της εκπαίδευσης. Η μελέτη υπογραμμίζει περαιτέρω ότι η χρήση συνόλων δεδομένων με υψηλότερη δομή σε αντίθεση με τα μη δομημένα σύνολα δεδομένων βελτιώνει την ακρίβεια πρόβλεψης και μειώνει τους κινδύνους θορύβου και υπερπροσαρμογής.

Συμπεράσματα: Τα αποτελέσματα υποδεικνύουν ότι τα μοντέλα που βασίζονται σε NLP μπορούν να προταθούν ως αξιόπιστη εναλλακτική λύση για την εκτίμηση/πρόβλεψη της διατροφικής σύνθεσης ενός τροφίμου από την λίστα συστατικών του. Αυτό προτείνει την επιλογή κλιμακούμενων και οικονομικά αποδοτικών εναλλακτικών λύσεων με βάση την ΤΝ σε σχέση με τις παραδοσιακές μεθόδους που βασίζονται σε εργαστήρια. Μελλοντικές ερευνητικές ανάγκες εντοπίζονται στους τομείς της βελτίωσης των δυνατοτήτων πρόβλεψης σε πραγματικό χρόνο, της βελτιστοποίησης των τεχνικών επιλογής χαρακτηριστικών και τελικά της χρηστικότητας αυτών των τεχνικών εντός των κανονιστικών πλαισίων της επισήμανσης τροφίμων. Η μελέτη αναδεικνύει τις δυνατότητες της μηχανικής μάθησης και της ευφυούς πρόβλεψης της σύνθεσης των τροφίμων για τη βιομηχανία τροφίμων ως εργαλείο αύξησης της εμπιστοσύνης των καταναλωτών προς τα συσκευασμένα τρόφιμα και την υποστήριξη της επισήμανσης υψηλής ποιότητας.

Table of Contents

List of Figures

Abbreviation Table

Abbreviation	Full Form
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
MAE	Mean Absolute Error
MSE	Mean Squared Error
R^2	Coefficient of Determination
SGD	Stochastic Gradient Descent
NLP	Natural Language Processing
LSTM	Long Short-Term Memory
ReLU	Rectified Linear Unit

Table of Basic Terminology

Term	Definition
Nutritional Composition	Breakdown of a food product's nutrient content and values.
Ingredient List	Ordered list of components used in a food product.
Machine Learning	Algorithms learning patterns from data to make predictions.
Natural Language Processing (NLP)	AI-driven analysis of human language for meaning extraction.
Food Labelling	Regulatory information displayed on packaged food products.
Data Preprocessing	Cleaning and structuring raw data for model input.
Model Training	Adjusting model parameters using data to improve predictions.
Loss Function	Measures prediction error, guiding model optimization.
Feature Engineering	Transforming raw data into meaningful predictive inputs.
Generalization	Model's ability to perform well on unseen data.
Overfitting	The model memorizes training data, reducing real-world accuracy.
Regulatory Compliance	Adhering to laws governing food production and labelling.

Chapter 1
Introduction

The years following the industrial revolution have been marked by a shift in the global burden of disease from communicable diseases to a dominance of non-communicable diseases like obesity, cancer and heart disease (Teng et al., 2023; Mohebi et al., 2022; Chew et al., 2023). The modern causes of morbidity remain primarily environmental with overnutrition, increasing dependence on processed foods, and undernutrition playing a role in the global double burden of disease (Popkin and Ng, 2022; Mozaffarian et al., 2021; Hill et al., 2022). Lately, the consumption of ultra-processed and packaged food has attracted the attention of the scientific community in terms of its adverse role in health. Epidemiological evidence provide links of ultra-processed food consumption with various health endpoints like mental health (Lane et al., 2022), overweight and obesity (Pagliai et al., 2021; Moradi et al., 2022; Rauber et al., 2021), type 2 diabetes (Levy et al., 2021), cardiovascular diseases (Juul et al., 2021a,2021b) and low-grade inflammation (Tristan Asensi et al., 2023) all linked to a general increased all-cause mortality rate (Suksatan et al., 2021) and decreased quality of life (Hosseinpour-Niazi et al., 2024; Rodríguez et al., 2022; Harris, 2024). In this context, nutritional interventions that enable healthier food choices are at the heart of public health globally, as almost every nation is struggling with population aging and chronic diseases stress the health sector. Policymakers are turning to the food industry with requests for higher transparency and accountability on the nutritional quality of their products and a more concise and actionable provision of nutritional composition data (Paulionis, 2008). The provision of clear, transparent and actionable nutritional composition data is a key request to the food industry as this allows for better consumer choices, simplifies regulatory actions and rectifies an increasing mistrust towards the food production system. Food labelling is instrumental in this aspect as the data provided on pack has the capacity to inform consumers, policymakers and any other stakeholder about the practices and processes involved in the production and distribution of a food and hence it allows informed decision-making and creates an element of trust and empowerment (Bacarella et al., 2015).

The provision of standardized, uniform food labelling creates a level playing field for allproducers, who are able to report their compliance with regulatory requirements in terms of

safety, appropriate use of ingredients and additives, as well as the use of appropriate food marketing strategies away from misleading claims about a food's role on health and wellness. On a larger scale, the data provided public on pack can serve as a live database capable of describing market trends, identifying regulatory gaps and needs, and performing risk assessment for health and safety matters (Facioni et al., 2020). Despite its potential to act as an open access database, the curation and collection of food labelling data poses significant challenges. The lack of uniformity in the national and regional labelling legislation, the differences in format, language and even the nature of mandatory data declared on the pack are some of the elements that introduce errors and bias when data from different regions are compiled in a single database. The differences in the legal requirements for the methods considered acceptable for regulatory use to generate and verify nutritional information introduce further challenges in terms of consistency and perceived accuracy of the data (Temple, 2020; Nayak & Waterson, 2019). Traditional methods to generate and verify a food's nutritional composition relay on chemical analyses, processes that although accurate are time-consuming and costly (Henderikx, 2017). When scalability is considered, linguistic diversity and variability of food product descriptions across regions and markets further complicate efforts towards the production of big food data structures. While the continuous introduction of innovations or renovations (new or modified food products) introduce the need for continuous updates of such datasets (Kasapila and Shaarani, 2011). On the flip side of this argument, the dynamic nature of food production, especially on an industrial level, requires structures that can track and identify the introduction of new ingredients, processes and recipes and their impact on a food's composition, tasks that are key for regulators and the public to assess the safety on food technology innovations (Van den Wijngaart, 2002).

Consumers are also increasingly becoming heavier users of food labelling data as they grow more health-conscious, and they search for food choices that fit their dietary restrictions, as in the case of low sodium or low-fat diets. In this pursuit consumers use both front and back of pack data, namely nutrition claims, nutrition facts tables and ingredient lists (Gholizadeh-Moghaddam et al., 2023; Mente et al., 2021; Delgado-Lista et al., 2022). Regulatory agencies worldwide, such as the Food and Drug Administration (FDA) in the United States, the European Food Safety

Authority (EFSA), the Australian Competition and Consumer Commission, and Health Canada, have long now established guidelines for food labelling (Díaz et al., 2020). These guidelines include the provision of both standardized nutritional composition data and ingredients lists; however, this is not a standard practice globally. On the contrary, ingredients list provision is a common practice across the globe and a regulatory requirement that is not met with increased costs for food producers to include on pack as it requires no additional analyses or investment beyond the food recipe.

Using a food's ingredient list to predict its nutritional composition could be proposed as an innovative way to promote low-cost food labelling prediction and support food producers, consumers and regulators in early nutritional assessment in environment without mandatory nutritional composition requirements. Natural Language Processing (NLP), a branch of artificial intelligence, offers significant promise in addressing these challenges. NLP enables computer systems to understand, interpret, and generate human language, such as ingredients lists and link it with numerical outputs such as nutrient contents (Chowdhary & Chowdhary, 2020).

Currently, an increasing volume of databases that provide both ingredients lists and nutritional composition data are available, some even open access, however only a fraction of solutions can leverage this data effectively, highlighting the need for intelligent language models capable of decoding and structuring this information. Deploying NLP algorithms can revolutionize the way foods are designed, and consumers and regulators interact with foods, since they enable the automatic prediction of nutritional composition from widely and easily available data on the ingredients used in a food's production recipe. NLP algorithms can use information readily available to food producers to help them consider whether a recipe fulfills a food's requirement for perceived healthiness and/or its role on diets with specific restrictions/preferences, and it can also help consumers and regulators receive acceptable predictions of a food's nutritional composition when the latter is absent. By implementing NLP-based systems, stakeholders can promote transparency in food labelling, remove barriers in information provision linked to cost structures and can promote a global culture of promoting access to nutritional composition declaration independently of local regulation.

Chapter 2
Literature review

This units presents existing food labelling legislation in the United States of America (USA) and the European Union (EU), with a focus on the nutritional component of the legislation. The use of Natural Language Processing (NLP) in the automation of label analysis is discussed alongside its practical applications and research gaps to date.

2.1 Food labelling legislation in the EU and the USA

The EU has always been considered one of the regions with strict control over food labelling. With multiple amendments from 1978 onwards, the food labelling legislation in EU has aimed to "*serve the interests of the internal market by simplifying the law, ensuring legal certainty and reducing administrative burden, and benefit citizens by requiring clear, comprehensible and legible labelling of foods*." The first mention of a regulatory requirement for pre-packaged foods to display ingredients lists and a nutritional declaration for energy, protein, carbohydrate, fat, fiber, sodium and vitamins and minerals levels can be found as early as 1990 under the Directive 90/496/EE, expanded in 2000 by the Directive 2000/13/EC, which not only set regulatory requirements for back-of-pack labelling but covered any nutritional claim made in the front-of-pack. Despite these and many other legislation, nutritional declaration was not mandatory for all foodstuff sold in EU, until 2011 when the EU Commission released Regulation 1169/2011 combining and updating all regulatory guidance on the provision of food information to consumers and making the use of standardized nutritional declarations compulsory for all foods sold in the EU (Berryman, 2014).

Based on this regulation every prepackaged food sold in the EU must declare the following information on pack:

- The food's name in a manner that is clear and transparent
- The net quantity of the food
- Ingredients list- except for fresh and unprocessed foods- which includes ingredients, additives and processing aids

- Nutrition declaration
- Allergen information highlighted in bold for clarity and readability
- Minimum durability in the form of “use by” date
- Name, address and country of origin of manufacturer
- Storage conditions
- Instructions of use
- Alcohol content, where relevant

The regulation was made mandatory for all prepackaged foods with specific provisions for minimally processed foods like fresh fruit, vegetables, meat, eggs, fish etc. It also expanded its use to foods sold in restaurants and canteens in terms of mandatory allergen declaration. To verify its implementation the EU Commission declared that failure to comply with the improvement notice was deemed a criminal offense.

In the USA, regulations on food labelling have evolved in similar but distinct manner. Despite overall similarities between the USA and EU in terms of the mandatory elements of food labelling, the technical aspects of labelling remain different. In the USA, the United States Department of Agriculture (USDA) and the US Food and Drug Administration (FDA) own the majority of food labelling legislations, with the former heavily involved in any agricultural produce especially poultry, meat, and eggs and the latter involved in the labelling legislation of the majority of the food supply in the USA. The main difference between the two being that the USDA requires pre-approval of all labeled data while the FDA does not. In terms of key differences with the EU, the USA system exerts higher control on the data provided on the nutrition facts panel (nutritional declaration) and it is more lenient when it comes with the expression of nutrition and health claims in the front-of-pack compared to the EU.

In the remit, of this thesis, the back of pack information are of higher importance and hence it is important to highlight key differences. The Nutrition Facts panel in the USA, contains the same nutrients as is in the EU with the addition of values for added sugar content, saturated and trans fats, and cholesterol. In terms of micronutrients, in the USA the declaration of calcium, iron,

vitamin D and potassium content is mandatory for all foods. Another key difference with the EU, is the fact that all values in the USA are expressed as content per serving size with serving sizes for each food category being regulated by the FDA. All manufacturers have to declare nutritional composition per regulated serving and not per 100 g of product as in the EU. Finally, an important element of standardization is introduced in the USA, which requires all nutrition facts tables to be black and white with a given format in terms of font size, dimensions, order of nutrient appearance and location on pack which significantly improves the ability to collect uniform data from food packages, either manually or with optical character recognition technologies (Berryman, 2014).

2.2 Natural Language Processing

Natural language processing (NLP) is the discipline of computer science that develops methodologies for computers to process (understand and interpret) as well as generate human language. As a field it shares methods with various disciplines including human linguistic, computation linguistic, statistical engineering, machine learning, data mining and human voice processing recognition and synthesis among others. First introduced in the 17th century by philosopher and mathematician Gottfried Wilhelm Leibniz (1646–1716) and polymath René Descartes (1596–1650), it formed the basis for the development of the language translation engine (Santilal 2020). The first documented, invention in the field of machine translation can be traced back to inventor and engineer Georges Artsrouni in 1933. Nonetheless, the work of Sir Alan Turing in the peri-World War II era and published in 1950 under the title Computing Machinery and Intelligence led to the creation of the famous Turing test, an evaluation criterion for machine intelligence (Turing 1936, 1950). Although at the time of the Turing test development, NLP research was mainly focused on language translation, NLP can be broadly described as the automatic or semi-automatic processing of human language (Eisenstein 2019). As a field it combines knowledge of linguistics and the mathematical concepts of logical theory into a new field this of computational linguistics, using tools from philosophy, cognitive science and agent ontology. NLP is central to human-computer interaction as it allows machines to analyze and interpret human speech (in various forms) and perform a number of tasks. NLP can be divided in three distinct components i) Natural Language Understanding (NLU), ii) Knowledge

Acquisition and Inferencing (KAI), and iii) Natural Language Generation (NLG). The main use of NLP is information extraction and/or retrieval. Information extraction refers to tasks related to the extraction of key information from written or oral sample of human language in an automated manner (Hemdev 2009). The task can be performed in structural or semi-structural documents or files which are machine-readable and it can be performed in multiple languages. Information retrieval on the other hand, refers to the organization, retrieval, storing and evaluation of information from documents, files, source repositories. It is more often used in textual information but it can be applied on multimedia such as video and audio knowledge bases. As such through NLP, machines can perform a series of information extraction and retrieval operations like automatic text summarization, data mining, sentiment and speech recognition analysis and even more advanced tasks like deep learning, and machine translation agent ontologies generation.

2.3 Large Language Models

Large Language Models (LLMs), a subset of NLP, specifically designed to assist AI applications that aim to understand, generate and manipulate human language. As indicated in their name (large) these models are typically characterized by their training in vast amounts of data, which improves their capacity to provide with results that are coherent and context-appropriate. Of course, they are limited to the predictions based on the input they are trained on, but in general existing LLMs like OpenAI's GPT series, Google's BERT, and Meta's LLaMA have shown unprecedented capabilities in translation, summarization, sentiment analysis and dialogue style question answering (Kirchenbauer et al., 2023).

These capabilities of LLMs are mainly rooted in the utilization of transformer models, a set of deep learning architectures that improve the efficiency of language processing and contextualization. Transformers use models to attribute weights to different words relative to their importance in a sentence, a mechanism called self-attention, which allows to better describe and mimic the nuances and complexities of human language. These processes, do not only allow for better language processing but it also elevates the ability to generate language that mimics human language. In the above-mentioned examples, LLMs have been shown effective in summarization, translation, query resolution but also creative tasks like story writing

and code generation with better performance both in understanding and language production as their training is expanded on larger datasets (Kirchenbauer et al., 2023).

The transformer-based models are usually built on an encoder-decoder structure. The BERT model (Bidirectional Encoder Representations from Transformers) an exception of this rule has proven highly efficient in understanding and generating human language. Unlike other transformers, the BERT model, utilizes an encoder-only structure and instead of processing text in oner direction (left-to-right, or right-to-left) mimicking human reading, it utilizes a bidirectional approach meaning that it analyses both the words on the left and right of any given word. This bidirectional approach allows BERT to collect data on the context to allows for more nuanced language understanding as the meaning of the same word can be different when used in different contexts. The model is built in two key phases, that of the pre-training building on extensive data to generate general-purpose contextual embeddings following by a fine-tuning stage which adapts the general embeddings to task specific context (Tsai et al., 2019).

In order to address BERT's shortcoming, the RoBERTa model was proposed in 2018. RoBERTa follows the same principles as BERT but it employs larger batch sizes, faster learning rates and it also has abolished the next-sentence prediction task of BERT, leading to improved model performance and better model training. RoBERTa also introduces novelties to BERT, like dynamic masking to improve general pattern learning, sentence packing for improved processing and byte-level BPE tokenization which allows for understanding of abbreviations and everyday language (errors or intended abbreviated/shortened use). The combination of improvements on BERT and the employment of those novelties allows RoBERTa to more efficiently process large datasets and provide with improved results in various NLP applications (Delobelle et al., 2020).

Google's Pathways Language Model (PaLM) is another interesting LLM example. With multiple different iterations PaLM is developed for use with scientific and medical applications (especially Med-PaLM2). The main tasks supported by PaLM are content analysis, summarization, reasoning, translation and code-generation and its novelty relies on the use of Google's Pathways machine learning system combined with few-shot learning capabilities, meaning it can be trained on minimally labelled examples. These allow PaLM to demonstrate versatility in applications especially in tasks that require pattern and statistical relationship identification which are the key

components of its design and the contributors to its high performance in a variety of domains (Zambrano Chaves et al., 2024).

Aside from language generation LLMs are heavily used in language processing. One such task is entity recognition. Entity recognition, the ability to identify and classify information within text is a key step in coreference resolution, information extraction and text categorization. In comparison to traditional techniques based on rule-based entity recognition the use of LLMs allows entity recognition in multiple languages and domains as they are capable to capture contextual nuances and transfer their learning from a domain or a language to another. As such LLMs demonstrate higher adaptability, accuracy, and efficiency compared to their traditional counterparts (Yadav and Bethard, 2019).

Similarly to entity recognition, LLMs are heavily used in text parsing. Text parsing, the breakdown of sentences in grammatic components, allows to understand syntactic structures and perform tasks like part-of-speech tagging and detecting sentence boundaries. Parsing can be used for better text comprehension but also for the generation of text in a more human friendly format (Tai et al., 2024). Parsing can also be used by LLMs to deliver structured responses like completed tables or forms or other tailored output formats. LangChain output parsing and OpenAI function calling are two popular output parsing techniques. LangChain's output parsing allows for the output to be customized and delivered in user predefined schemas and formats, while OpenAI's function calling uses a simpler dictionary-based approach (Liu & M'hiri, 2024).

2.4 Named Entity Recognition (NER)

Named Entity Recognition (NER), the task of identifying and classifying entities found in text into predefined categories, is central in NLP. NER helps machines understand the role of words in a sentence like whether they indicate names of people, organizations, locations, dates or other predefined significant terms. NER helps transform unstructured text to structured information and through it, applications like information retrieval, question answering and machine translation can be developed.

In its early development, NER required the creation of predefined rules and dictionaries of known entities, gazetteers. These allowed for reasonable accuracy in specific domains with substantial rules and dictionaries. Modern NLP require larger NER capabilities to ensure

appropriate understanding and meaning information extraction, making these labor-intensive early NER approaches unscalable in todays AI industry.

To overcome these issues, probabilistic models like the Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) were introduced. These models use annotated data at a training stage, successfully reducing the need for manual feature engineering (Lample et al., 2016). These advances though still faced limitation when machine learning was taking place in multiple languages or when specialist texts were analyzed. These limitations were linked to the ability to understand context and semantics. In the 2010s, deep learning revolutionized NER through the introduction of Neural Networks and their impact on identifying hierarchical representations of text. Techniques like the Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) architectures allowed for NLP to learn through context and sequences analyzing both the structure of the sentence and the neighboring text around every word. The combination of deep learning BiLSTM with probabilistic models like CRF showed increased NER performance indicating that word-level embeddings and character-level features can be co-analyzed for better performance (Lample et al., 2016).

LLMs like BERT, further help improve NER by allowing for the processing of entire sentences in parallel. BERT and its successors, RoBERTa and GPT, can achieve state-of-the-art performance in NER tasks as they combine unsupervised pretraining on a large array of texts followed by fine-tuning on specialized context specific datasets (Kenton et al., 2019). This ability of BERT, makes it especially effective in understanding context and identifying relevant entities in complex, long and ambiguous sentences. However BERT continues to underperform in domain specific texts, where the use of language is different to that of the everyday written word or literature. Similar to Med-PaLM2 and the need for LLMs for specialized texts, scientists have developed domain specific NER systems and training datasets, like the BioBERT and FinBERT which have improved NER accuracy in medical and financial text analysis, respectively (Lee et al., 2020). In this context, NER is trained on contracts, legislations and policy documents to identify and extract data like clauses, parties involved and identifying risks of non-compliance. In the case of BioBERT, tasks are tailored to healthcare needs, like extraction of patient information, diagnoses and treatments from medical history records.

At this point, it is important to highlight that the majority of LLMs and NER architectures are developed and trained on English language data and are hence prone to underperformance in multilingual environments with differences in syntax and sentence architecture. Models like mBERT and XLM-R (Cross-lingual Language Model with RoBERTa) were developed following a pre-training step on texts from multiple languages which allowed then to perform cross-lingual transfers and cross-lingual generalisation even when having access to limited training data (Conneau et al., 2020). In the area of linguistic diversity, the application of NER on social media generated texts has its own challenges. The use of language in social media presents with informal tone, abbreviations, slang, colloquial phrases, the use of everyday words but with a completely different meaning and often with misspelling, unintended or even intended for character saving purposes. Alongside cross-lingual models, research applies contextual embeddings and noise-robust training methods to adapt NER capabilities for social media language use (Akbik et al., 2019). Especially in the area of novel words or word usage especially prevalent in social media, NER is still in need of continuous learning mechanisms and the update or entity lists in order for technology to become familiarised with newfound terms or newfound uses of existing terms (e.g. differentiating apple the fruit from the tech company) (Bhowmik, 2021). On the other hand, NER is also prone to propagating societal biases linked to the use of words like gender and ethnical discrimination. As the models are trained on historical data, they are likely to pick up nuances that are considered discriminatory nowadays but the modern use of these words in the modern context is relatively less frequent. Approaches like data curation, algorithmic fairness measures and evaluation practices to promote transparency might be beneficial in addressing such issues (Buolamwini and Gebru, 2018).

Overall, NER is in need of methods that allow entity linking potentially in knowledge graphs which could allow for disambiguation. The identification of the location of an entity on the nodes of a knowledge graph could allow to reduce ambiguities by better contextualization and the association of additional attributes and connections to (Bhowmil, 2012). However, as NER methods expand so are the needs for few-shot and zero-shot learning techniques which can be applied on minimally labelled data and can improve technology access in low-resource environments and non-specialist domains.

2.5 Applications in food labelling

NPL utilization tailored towards the food industry is a relatively new field with a few examples primarily on ingredient extraction and customer feedback management. In theory, NLP could be used to process the large volumes of textual data produced by the food industry in the form of ingredient lists, marketing claims and other claims or information provided on pack. The application could both product development facing or aiming to reduce processing times for regulatory compliance testing. Through text processing such as ingredients lists, NLP, can allow for ingredient calculation, allergen detection or even regulatory compliance checks in terms of format, wording etc. Given the substantial manual effort associated with these tasks their automation can not only reduce time and resource requirements but it can also reduce human error and provide higher quality output (Hu et al., 2023).

A key requirement in the field of food science linked to food labels in the capacity of stakeholders, mainly consumers but also producers, to automate tasks like food classification from their name into food groups, estimation of their nutritional composition and even portion estimation. Tools like the goFOOD™ (Lu et al., 2020), FoodSky (Zhou et al., 2024) and NutriBench (Hua et al., 2024) have been built with NLP capabilities to test the potential automation of such tasks and their relevant accuracy. The existing solution leverage NER capabilities in handling ingredient lists and marketing claims allowing them to extract specific ingredients or even identify nutrition and health claims through sentiment analysis. In the overall field of food and nutrition, NER methods have also shown potential in recipe analysis from digital sources and dietary recommendation identification. Much like in other domain, early innovations like FoodIE and drNER relied of predefined rules for food entity recognition, but it is unknown how deep or machine learning techniques like BERT, and BiLSTM-CRF could improve entity extraction precision (Popovski et al., 2020; Wei et al., 2019). These methods will need to consider the specialized nature of the text found on labels and the domain specific use of the language when food science is concern, in order to act as domain specific application capable of providing high quality data while reducing cost and time commitments (Miyazawa et al., 2022).

In the pursuit of improved and more extensive application of NLP technologies in the food sector, especially the packaged food sector, there are a number of domain-specific challenges to be considered and addressed (Ma et al., 2024; Holland et al., 2020):

- Data Availability and Quality: Food data are either small in scale to train NLP models but provide with high quality data or large with multiple entries but questionable data quality and architecture. Only a few examples exist of expert curated datasets with sizes suitable for NLP training and those are often proprietary.
- Domain-Specific Language and Terminology: The terminology used on food labels is often a mix of regulatory definitions, chemical terms, abbreviations and generally specialized text which is only relevant in the food label context and cannot be generalized in other domains, e.g. low-fat, reduced fat, plant-based fat, organic, natural. Fine-tuning could require more than one dataset and even the combination of datasets from different domains such food science and policy documents.
- Multilingual and Multicultural Diversity: Food labels exist always in the local language and the presence of the same information in dual languages (one being English) is not necessarily common across the globe. Even when dual language labelling exists, only a part of the labelling is translated and the quality of the translation can be questionable. At the same time, due to regulatory differences the same term might be describing a different characteristic in different countries e.g. a high protein food provides with 10g of protein per serving while in the EU a high protein food must provide 40% of the food's energy from protein-meaning that it must contain at least 0.1g of protein for every calorie it provides.
- Linguistic diversity: Food product and ingredient descriptions vary across countries with colloquial names often being used on food labels without necessarily the concurrent use of the food's or ingredient's standardized name in a botanical or chemical format.
- Complexity of Food Products: As ingredient list are non-weighted and are simply a list of ingredients in declining order, their use for nutritional composition estimation is hindered by the complexity of ingredient interaction, processing methods and food format especially its water content (a common factor not present in any food label).

2.6 Ethical considerations in the use of NPL in Food Label Analysis

Although NLP has the potential to improve the mining and utilization of food data by consumers and the food industry it needs to be treated with cautious until issues around fairness, accuracy and transparency. These challenges can be traced back to algorithmic biases, privacy concerns and accountability, all central to maintaining user trust in AI technologies.

The primary concern is linked to the accuracy of NLP algorithms. In the context of prediction of labelled information, accuracy is not just a technical attribute of the algorithm but an element of legal and ethical concern. Inaccurate prediction of labelled information, can expose food producers to regulatory misalignment with legal consequences and equally it can expose consumers to misleading choices. The existing scepticism related to the "black-box" algorithms employed by AI, would be severely influenced if a manufacturer was found charged for declaring inaccurate information or if a consumer used an AI generated nutritional composition prediction that misevaluated the sodium, sugar, allergen etc content of a food and impacted their dietary choices even potentially their health.

This highlights elements of accountability linked to the automated generation of health and regulation sensitive information. At this point, determining responsibility in terms of ownership of AI generated erroneous data is complex and it is unclear whether it falls on the developers, the deploying companies or the regulator for not foreseeing relevant clauses for AI use (Jacobs et al., 2021). This lack of responsibility although it does not impact research directly it hinders innovation as the framework of application is unclear for such technologies and so any resource saving potential they might document remains non-actionable by the end-users, which in the case of the industry is risk averting.

The second major concern is linked to algorithmic biases. The use of non-representative or lower quality data to train NLP systems is the main source of algorithmic biases. Training on such data could lead to disproportionately favouring specific outcomes when it comes to predicting values. For example, an AI model trained on data from a specific country or manufacturer is more likely to predict values within the range provided during the training even if the actual data for a food from another country or manufacturer are indicative of a better nutritional profile. Buolamwini and Gebru (2018) have shown this concept of bias to be true with AI perpetuating

discriminatory behavior against marginalized group mimicking the training data and the even raise issues around the difficulties around rectifying or reprogramming an NLP to identify and address such biases. Martin (2019) highlights that such biases should be resolved with the development of what is called ethical AI frameworks which ensure inclusivity. In the case of food labelling, these issues could be resolved with the provision of better training data, higher versatility and representative, but this in turn highlights the need for larger scale open access data on food labelling.

This is linked to general concerns over General Data Protection Regulation (GDPR) and overall privacy issues. Although food labels exist in the public domain, data ownership is unclear, meaning that according to legislation erroneous data on the label are attributed to the manufacturer but it is unclear what happens when erroneous data are found on food label repositories. Similar to private data, food data are company attributable and any conclusions, concerns or misuse could harm unfairly the interests of specific manufacturers. If food label data are crowdsourced, a common approach, then erroneous data can be attributed to specific compilers and any legal actions might lead to privacy infringements. In this context the volume of training data needed for the development of AI solutions can be hindered by data ownership and privacy concerns and equally the use of AI generated data suffers from unclear legislation on the topic. The European Commission's guidelines for trustworthy AI (2019) is a suitable starting point to address these concerns but in this context food data should be collected and curated under human oversight, all solutions should offer transparency and they should avoid any discrimination. To-date, AI innovation in food data is limited and there is much to be learned about how the application of these principles may be hindered in the real world.

2.7 Identified research gaps

The arguments presented so far, highlight the potential of AI applications to revolutionize the food and nutrition sector and to introduce a traditional sector to the 4th industrial revolution. The management of text from food labels in order to predict key food characteristics has multiple potential applications but research in this direction needs to consider and address to any degree possible potential roadblocks and challenges. Figure 1 shows the main research gaps identified in the use of NLPs for food label analysis. The key areas identified include ingredient recognition

when technical or local terms are used, the parsing of text and subsequent translation to allow for understanding and overcome language barriers, the handling and identification of data errors and even the recognition of cross regional regulatory compliance issues and their proposed solutions by tailored LLMs.

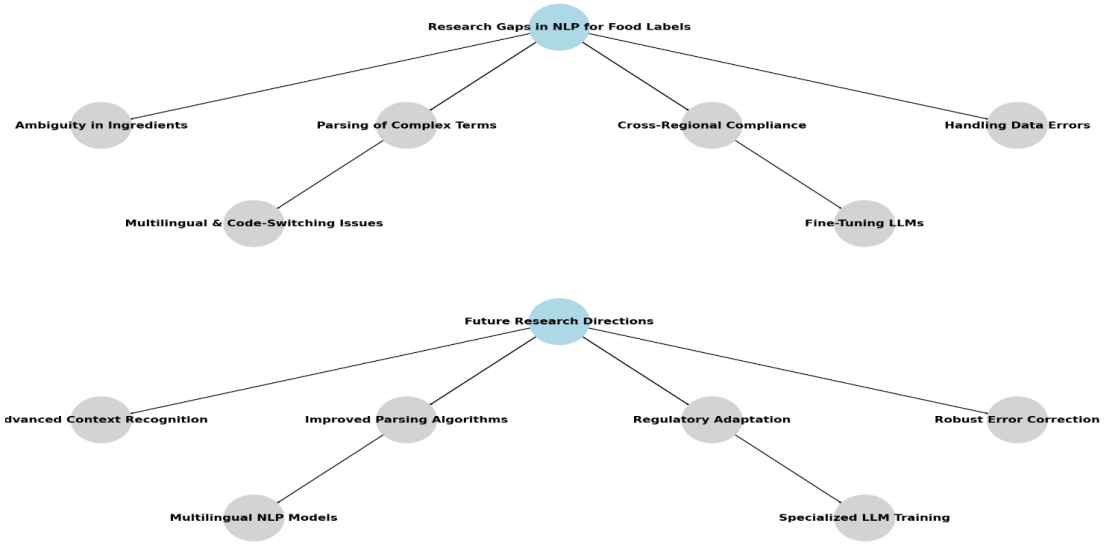

Figure 1 Research Gaps and Future Directions In NLP

The main bottleneck in the development and application of NLPs on food label data is linked to the nature of the data themselves. Food labelling often includes nuanced, expert terms combined with colloquial phrases or ambiguous use of words. Depending on the type of food data ingredients list might be too nuanced or even described using specific codes while other terms like marketing claims can be too ambiguous and subtle like understanding the difference between “natural” and “artificial” flavorings. The development of models designed to parse highly technical terms is in the early stages when it comes to recognizing and categorizing chemical structures/compounds and food ingredient terminologies. The lack of universal agreement on the recommended terminology used introduces linguistic variability which complicates the task and the lack of extensive datasets with available ingredients lists for a large volume of foods limits training capabilities.

Advanced parsing methodologies tailored for the application on food labels will be required. These methodologies will be asked to address issues linked to the different regulatory

requirement of food label information, in terms of format, wording, use of symbols, phrases even the use of abbreviations. Many countries refer to ingredients using their botanical name, others using their local name, the chemical name or even have predefined codes for additives with long names to allow for space saving on pack. In countries, ingredients are shown in declining order while in others ingredients that are themselves complex structures (e.g. breadcumps) are broken down to their own ingredients within the same ingredients list (in brackets next to the original ingredient used). Finally, NLPs will need to be able to perform these tasks in multiple languages as provision of nutritional data in English is not compulsory across the globe.

These challenges are linked to prediction accuracy and the usability of the proposed NLPs but they can be addressed using suitable datasets that provide large volumes of data in a structured manner and suitable for use an NLP input. However, such datasets are sparse. Food data is often limited in terms of food items included or when they have large numbers of food data then the data are either unstructured, they lack data quality checks and are prone to mistakes and empty values. As such, future research should identify NLPs that are capable of handling unstandardized formats and error prone datasets and evaluate their capacity to read and interpret textual data available on-pack. These activities will allow to test the capacity of existing LLMs in handling food data for specific applications, identify fine-tuning areas and propose the need for specialized vocabularies, abbreviations or even develop methods to handle these specialized contextual nuances. Through this process not only LLM research will be advanced but the general food data science field, currently under exponential growth, will receive crucial feedback for data specifications that could enable future applications.

2.8 Research Questions and Study Orientation

The main research aim of the current thesis' research was to examine the feasibility of using a food's ingredient list to accurately predict its nutritional composition, namely its protein, total fat, total sugars, dietary fiber and sodium content. To achieve that the study will employ language models to predict nutritional information via the extraction and processing of unstructured textual data (ingredients list).

This objective was split in three objectives around the main challenges foreseen in the algorithm development process:

1. How do different loss functions and other experimental conditions (epochs, batch size, learning rate etc.) influence the model's predictive accuracy for nutritional data?
2. How does training on category-specific datasets (e.g., cheese only) compare with generalized models trained on mixed categories in terms of validation loss and R^2?
3. Which nutrients are more accurately predicted, and which pose greater challenges for accurate prediction based on ingredient lists?

These three objectives describe a step-wise experimental setup covering the model set-up process from data preparation, selection to model training parameters. More specifically the planned comparative evaluation of optimization strategies will i) help assess the most effective loss function (among MSE[1], MAE[2], Huber, SmoothL1Loss) as well as other experimental conditions that provide improved validation and training loss, ii) evaluate the impact of data segmentation on model generalization and predictive performance and iii) identify domain-specific limitations and potential biases in learning patterns linked to the prediction of specific nutrients.

2.9 Innovation of the Study

This study aims to significantly contribute to the field of automations in food data science and especially in AI-assisted nutritional composition prediction. From a technological stand point the study will conduct a comparative analysis of multiple loss functions (MSE, MAE, Huber, SmoothL1Loss) and experimental conditions like epochs, batch size and learning rate to investigate their impact on model stability, generalization, and behavior in high-noise environments. This approach, goes beyond existing literature, such as the study by Ma et al. (2021), which is limited to traditional regression algorithms. Secondly, the study will investigate whether the variability expected within a diverse food composition dataset including foods from various food groups, will impact the training and validation performance of the algorithm. This reflects an important bridge between food science and data science as it is based on the

[1] MSE: Mean Squared Error

[2] MAE: Mean Absolute Error

knowledge that food composition data are likely to be more homogeneous within the same food category and it is currently unknown whether this homogeneity will impact prediction accuracy. To achieve that the study will be based on the largest dataset with both nutritional declaration and ingredient list data available which will allow to test the impact of using large, unfiltered dataset with multiple food categories for modelling training compared to applying preprocessing and data selection methodologies to reduce heterogeneity. Finally, the research will go beyond classical statistical methods (R^2, scatter plots etc.) in evaluating a model's predictive accuracy per nutrient as it will employ Bland-Altman plots to analyze the agreement between predicted and actual values using a gold-standard approach.

The research will also provide indications on the training cost and convergence time of the proposed models, which will allow for better translation of the theoretical machine learning tools to real-world applications.

Table 1 Key Innovations and Original Contributions of the Study

Innovation Area	**Contribution**
Loss Function Evaluation	Testing the role of the MSE, MAE, Huber, and SmoothL1Loss loss functions on model performance
Dataset Structuring	Testing the impact of using food category specific datasets on model performance compared to unfiltered datasets
Model Interpretation Tools	Use of advanced statistical methods, Blant-Altman plots, to measure agreement between predicted and actual values.
Computational Efficiency Assessment	Indications of training costs and associated predictive improvements per architecture
Application Scope	Performance of nutrient specific prediction accuracy
Integrated Methodological Strategy	Co-evaluation of deep learning and data structure parameters on model performance

Chapter 3
Methodology and Experimentation

This section covers the methodology for predicting the nutritional composition of foods using Natural Language Processing (NLP) and deep learning techniques. It includes the dataset used, data preparation, model structure, training process, optimization techniques, and evaluation metrics. The experimental procedure examines differences in performance across dataset size and machine learning algorithms, comparing regression and classification methods. The effect of dataset specificity on prediction accuracy is also examined, as well as the use of different cost functions and optimizers. The combination of text-based representations and numerical nutrient prediction is emphasized to implement transformer-type integrations on food labels.

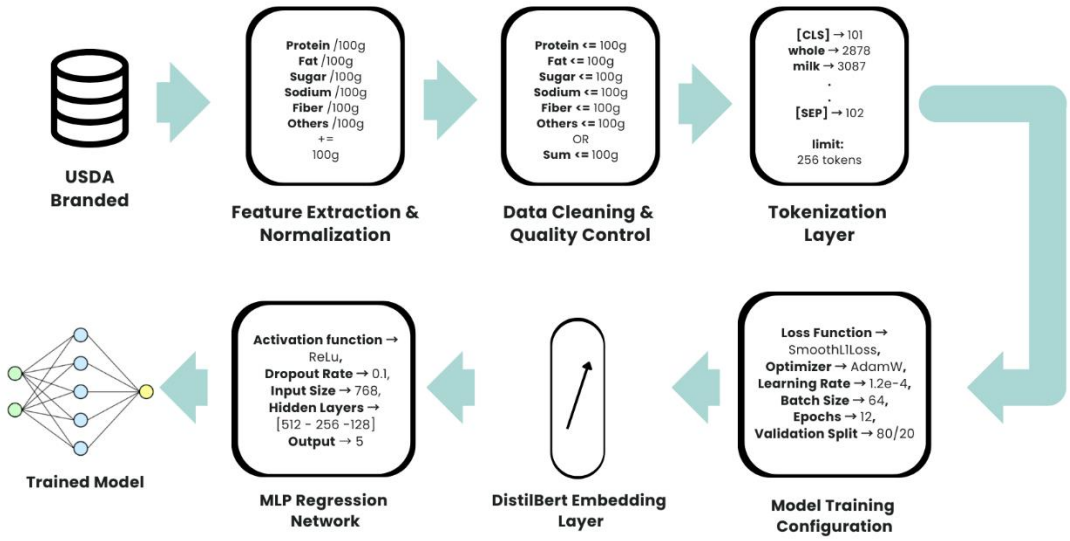

Figure 2: Methodology Architecture

3.1 Dataset and Subjects

The dataset is sourced from the USDA FoodData Central database and focuses on branded food products, with the goal of obtaining a complete list of ingredients and nutritional values. This is a critical choice for building a reliable model, as it covers a wide range of processed and unprocessed foods—reflecting real-world eating habits.

Information such as the Food ID, ingredient list, and nutritional data are clearly structured, making it easy to match text ingredients with nutritional values. This organized format also supports the use of natural language techniques to parse ingredient lists, while combining this information with numbers into a single, integrated system.

The provided dataset already standardizes nutritional values per 100 grams, while the innovation of this study adds an extra step, calculating values in grams (g), thus ensuring fair comparisons between different foods.

Normalizing values to a single scale and unit of measurement helps the model not only ensure comparability, but also align with regulatory and scientific standards of nutritional analysis, as most food labels and dietary guidelines are based on the per 100g format.

Table 2 Standardized Nutrient Table (per 100g)

Nutrient	Role in Diet	Standardization (per 100g)
Protein	Essential for muscle synthesis and cellular repair	Measured in grams
Total Fat	Provides energy and supports cell function but must be consumed in moderation	Total fat in grams as the sum of all fatty acid subcategories
Sodium	Regulates fluid balance and nerve function; excessive intake linked to hypertension	Converted to g from mg to allow for same unit of reference
Total Sugar	Primary source of energy often added to food products during processing	The sum of mono- and di-saccharides in g (a subsample of total carbohydrates)
Fiber	Supports digestion and helps regulate blood sugar levels	Expressed in grams, considering soluble and insoluble fibers

To achieve high resolution in the nutritional assessment of foods, the dataset includes five main categories of nutrients: Protein, Total Fat, Sodium, Total Sugars, and Dietary Fiber. These components were selected based on their importance in regulating metabolic functions, their contribution to the pathogenesis of chronic diseases, and their decisive effect on the overall nutritional balance.

Protein is a macronutrient of critical importance, as it participates in anabolic and catabolic processes, in the structural integrity of tissues and in enzymatic function. Total fats function as a basic energy substrate, while at the same time supporting fat-soluble vitamins and regulating hormonal and cellular functions. Sodium, as the main cation of the extracellular space and a basic electrolyte, participates in the maintenance of acid-base balance, fluid homeostasis and the transmission of nerve impulses. Excessive sodium intake is causally related to hypertension and cardiometabolic diseases. Total sugars include both endogenously existing sugars in foods and those added during industrial processing. Their measurement is necessary for the assessment of the glycemic load and the possible effect on glucose metabolism. Dietary fibers are complex carbohydrates that are not digested by the enzymes of the digestive tract. Their presence has a proven beneficial effect on intestinal function, glycemic regulation and lipid profiles, which makes it necessary to incorporate them into nutritional adequacy prediction models.

This research does not collect information about foods, but proposes a new way to predict their nutritional composition. Unlike previous studies that rely on general rules, laboratory analyses or data collected by hand, this one uses machine learning models that have been trained on large databases with reliable nutritional data. Thus, they manage to calculate nutritional values with accuracy.

The innovation is that ingredient lists are directly linked to the nutritional value of each product. This allows us to convert plain text (such as a food label) into structured data, suitable for analysis by artificial intelligence systems. This approach improves the accuracy of predictions, while at the same time opening the way to useful applications, such as automatic nutritional assessment, product labeling and compliance monitoring with nutritional regulations.

3.2 Experimental Design

The study's experimental design is based on the usage of natural language processing (NLP) with deep learning to predict the nutritional composition of food products based on ingredient lists. The goal is to build a predictive model that can analyze text-based ingredient data and accurately estimate macronutrient values. Unlike traditional approaches that rely on pre-defined nutrient look-up tables or statistical heuristics, this study introduces an end-to-end learning framework that directly maps textual descriptions to numerical outputs. The process begins with data preprocessing, where ingredient lists undergo tokenization, normalization, and feature engineering to ensure structured input representation. This transformation is essential, as ingredient nomenclature varies significantly across datasets, requiring an advanced text-embedding technique such as DistilBERT to capture linguistic and semantic nuances.

Table 3 Experimental Pipeline Overview

Stage	Description	Key Techniques
Data Preprocessing	Cleaning, tokenization, and transformation of ingredient lists into structured data	Tokenization, normalization, feature engineering
Model Architecture Definition	Development of deep learning models using NLP embeddings and regression techniques	DistilBERT embeddings, MLP regression, activation functions
Training	Training of models on structured datasets with varying hyperparameter settings	Batch processing, adaptive learning rates, dropout regularization
Evaluation	Assessment of model performance using error metrics and dataset-specific comparisons	MAE, MSE, R^2 score, dataset stratification, loss analysis

The Experimental Pipeline Overview table systematically presents the research process, detailing data preprocessing, model architecture, training, and evaluation, ensuring clarity in methodology while highlighting key techniques for each stage.

One of the key innovations in this experimental setup is the structured comparison between category-specific datasets and general datasets. Traditional food composition studies often suffer from high variability in ingredient formulations, making it difficult for a model to generalize across different product types. By segmenting the dataset into focused categories (e.g., cheese, beverages, snacks), the study aims to assess whether narrowing the scope of data improves predictive performance. The comparative table highlights the key trade-offs: category-specific models exhibit higher accuracy due to reduced variance in ingredient patterns, whereas general datasets offer broader applicability but at the cost of increased model complexity. This approach ensures that the research findings provide valuable insights for both domain-specific applications and large-scale food industry automation.

Table 4 Comparison of Category-Specific vs. General Datasets

Aspect	Category-Specific Dataset	General Dataset
Dataset Scope	Focuses on specific food types (e.g., cheese, beverages)	Includes diverse food categories, increasing generalizability
Data Complexity	Lower variability, as all samples belong to a single category	Higher complexity due to ingredient and nutrient diversity
Predictive Accuracy	Higher accuracy due to targeted learning	Lower accuracy, as the model must generalize across many food types
Computational Efficiency	More efficient training, requiring fewer epochs to reach convergence	Computationally demanding, requiring more extensive training

Table 4 presents the differences between two different data types, emphasizing the accuracy of predictions, the complexity, and the computational efficiency of machine learning models. The model is designed to be flexible and easy to adapt, so that each developer can “tune” it according to the characteristics of his own data set. It uses DistilBERT, a lightweight natural language

processing model, to convert ingredient lists into numerical vectors (embeddings). These vectors are then given as input to an MLP (multi-layer perceptron) network used for regression. In the framework of the experiments, various variations were tested, such as different numbers of hidden layers, activation functions, and regularization techniques.

To find the most effective activation function, a comparison was made between GELU, Swish and ReLU. The model training process also includes the use of the adaptive AdamW algorithm, which allows for adjusting the learning rate, making it more efficient. As a loss function, SmoothL1Loss was chosen after a comparative evaluation with MSELoss, L1Loss and HuberLoss. The results showed that SmoothL1Loss is more robust to the presence of extreme values (outliers), without losing the accuracy of the regression.

In addition to accuracy, computational efficiency also plays an important role in experiments. Models trained in specific food categories (e.g. cheese) need fewer training cycles (epochs) to "learn", probably because their data contain less noise. In contrary, general models need more computational resources, since they need to understand more complex patterns. Table 4 shows that general models have advantages in terms of scalability, but their performance can be reduced due to the large variety of features. This suggests that, in the future, a combinatorial training approach may be more effective: first we train a general model and then we adapt it to specific categories, to achieve a good balance between speed and performance.

3.3 Data Preprocessing

The data preprocessing stage is a vital part of this study, as it transforms raw ingredient lists into a format that deep learning models can effectively understand. Table 5 presents the preprocessing steps for the data, showcasing the importance of the methodology, purpose, and impact on the model's performance.

More specifically, in the data normalization and cleaning stage, unit inconsistencies were handled across the dataset. Specifically, Sodium values were reported in milligrams (mg) and were converted to grams (g) for consistency with the rest of the nutrient units. Moving on to the cleaning stage, where entries were filtered based on nutrient totals. Any sample where the total nutrient content—or a single nutrient—exceeded 100g per 100g was removed to keep the dataset realistic and standardized.

Moreover, after the data was normalized and cleaned, the ingredient list of the entry was fed to the DistilBERT tokenizer. In this step, text is broken down through tokenization. Limitations on the token length —256 tokens — and omitting entries of which their ingredient list exceeded the limitation, ensured noise reduction and improved the structure of the input. Duplicate entries can introduce biases in training, causing overrepresentation of specific food types. Tokenization breaks down ingredient lists into smaller, manageable units. This acts as the first step in converting text into numerical format so that the model interprets each item using NLP embeddings. This ensures that even complex ingredient formulations are accurately recognized and mapped to nutritional values.

Feature engineering introduces an innovative approach to handling missing nutritional data, which is often a limitation in real-world datasets. The inclusion of an 'Others' category ensures that the total composition remains at 100%, preventing negative or unrealistic predictions. This structured representation of untracked components enhances model stability and interpretability by providing a more comprehensive view of food composition. Additionally, the 'Others' column plays a crucial role in distinguishing foods with similar macronutrient ratios but varying ingredient compositions, improving the model's ability to differentiate between products with subtle nutritional differences.

Without this feature, the model would be forced to infer missing values arbitrarily, potentially reducing prediction reliability. By explicitly accounting for nutrients that are either unreported or indirectly derived, this method reduces the risk of misclassification and improves overall model generalization. This is particularly significant when dealing with processed food products that contain additive compounds, ensuring that the model can make accurate predictions even in cases where full ingredient breakdowns are not provided.

Table 5 Structured Data Preprocessing Framework: Techniques, Implementation, and Model Impact

Preprocessing Aspect	Implementation Strategy	Engineering Significance	Expected Outcome
Tokenization	Long ingredient lists exceeding the max token length (256) were omitted to prevent duplicate creation.	Improves computational efficiency while preventing data duplication.	Ensures faster training times, higher model precision.
Normalization & Cleaning	Conversion of sodium from mg to g. Entries that exceed 100g or sum to >100g were omitted.	Standardizes numerical values, preventing scaling inconsistencies.	Enhances stability in model predictions and convergence speed.
Feature Engineering	Creation of 'Others' category for untracked nutrient data.	Provides structured representation for missing values.	Reduces prediction errors related to incomplete datasets.

3.4 Text Representation & Model Architecture

This section explores how DistilBERT processes ingredient lists into embeddings, to act as an input in the regression network through a multi-layer perceptron architecture.

3.4.1 Text Processing with DistilBERT

The use of DistilBERT in text processing allows the conversion of ingredient lists into structured representations (embeddings) efficiently, based on deep contextual representations instead of simple individual terms. The transition from text to numerical representations is a high value methodology that establishes the semantic understanding of the ingredient composition. Specifically, by using the DistilBERT-base-uncased tokenizer for text processing, we manage to

convert the raw ingredient lists to a [CLS] token representation, which is then converted into embeddings. This approach significantly helps in generalization, even for rare or unusual ingredients that do not appear often in the dataset. Tokenization acts as the first step in converting ingredient lists into numerical data, allowing the model to understand each item through DistilBERT embeddings and accurately map complex formulations to their nutritional meanings.

In order to achieve unity in token length, a maximum limit of 256 tokens per entry was applied, providing consistency and preventing the model from being burdened with unnecessary information - while maintaining low computational time without a significant reduction in the size of the data - techniques such as padding, truncation and attitude masking were used. Padding is used to bring the shortest suggestions to the same length as the rest, adding padding tokens at the end. Attention Mask helps the model understand which tokens are essential and which act as padding, to take into account only valid data in the final output. In this study, instead of a simple truncation of the long lists, records exceeding the maximum length were excluded from the dataset. The reason is to avoid repetitive or overloaded lists that could degrade the quality of the input.

The model gains linguistic knowledge from large scale text by the usage of the pre-trained DistilBERT to generate embeddings. Enhancing this way, its generalization power when asked to predict nutrient values. This is a key difference of such methodologies when they are compared to the traditional statistical approaches. These self-attention mechanisms dynamically assign importance to different ingredients. This feature helps to deal with the problem of data sparsity. Through these preprocessing techniques, both efficiency in memory management and system stability improve, while limiting the effect of unnecessary or "noisy" data. The system avoids dependence on manually designed features, thereby allowing an automated, scalable and more customized approach to the analysis of food composition.

3.4.2 Multi-Layer Perceptron (MLP) for Regression

The Multi-Layer Perceptron (MLP) architecture is used to transform the 768-dimensional embeddings of DistilBERT into numerical predictions of nutritional values, using a fully connected neural network optimized for regression problems. The input data — derived from semantically

enriched ingredient representations — is structured numerical information that allows the model to apply nonlinear transformations, enhancing its ability to capture complex relationships between ingredients and nutritional properties.

The network includes configurable hidden layers — arranged in a 512 → 256 → 128 neuron format — allowing flexibility in choosing the complexity of the model. Through this gradual reduction of dimensionality, the feature extraction process is enhanced, while the overfitting is limited, and essential dependencies within the data are preserved. Each layer of the MLP optimizes the component representations by extracting hierarchical patterns that enhance the accuracy and ability to match the ingredient list description and their corresponding nutritional properties. This hierarchical structure improves both the generalization of the model to new data and the learning efficiency.

To enhance the training stability and prevent overfitting, regularization strategies and activation functions play a critical role. Dropout with a percentage of 0.1 is applied, introducing random neuron deactivations during training — which reduces the model's dependence on specific patterns and increases generalization to unseen data. In combination with the dropout rate, activation functions were tested — such as ReLU, Swish and GELU — with ReLU providing the most desired outcome since it guards the output of each layer to be non-negative.

Using these techniques significantly contributes to the stability of the model and enhances the accuracy of its predictions. Thus, when the model parses through an ingredient list, we can be confident that its nutritional estimates are reliable and accurate — even for different or more unusual food products.

3.5 Training and Optimization

In this section, we will display how the decisions were made regarding the choice of the loss function, optimization strategies, learning rate, and batch size. We will focus on understanding how they affect the stability and efficiency of model training.

3.5.1 Loss Functions

Loss function selection plays an important role in the training stability and mostly to the accuracy of the generated predictions of the model. To conclude on which loss function offers the most optimal results, the study tested SmoothL1Loss, MSELoss, L1Loss and HuberLoss. Each of the loss functions considered has its own advantages and disadvantages — some are more sensitive to small deviations, while others handle outliers better. For the purposes of this study, SmoothL1Loss was found to perform best. SmoothL1Loss seemed to be the most appropriate choice, because it helps the model learn at a steady and smooth pace, without "getting confused" or overreacting when encountering extreme or strange nutritional values. With MSELoss, large errors are overly magnified, making the training unstable. On the other hand, L1Loss can drive the model to the opposite extreme — not "caring" enough about large deviations. SmoothL1Loss offers a middle ground, with smooth transitions that make the overall learning process more stable and reliable. There are other options, such as HuberLoss, that can give similar results — but in practice requires a lot of experimental tuning to end up with a significant difference from the others. Overall, SmoothL1Loss was chosen because it is simple to implement, worked reliably in different training scenarios, and handled both specialized category sets and broader general data equally well.

By observing how the loss function behaved with different dataset sizes, we learned a lot about how quickly and smoothly the model learned. With smaller datasets, we noticed that the loss fluctuated much more — indicating that special care is needed to avoid overfitting. Out of the functions tested, SmoothL1Loss handled the situation optimally, both on small datasets — 10,000 entries — and medium-sized datasets — 80,000 entries — maintaining both train and validation losses low and displayed consistent behavior throughout the experiment. When it comes to MSELoss and L1Loss, they produced unstable loss curves, with sharp fluctuations making it difficult to track the learning. When the dataset size exceeded 150,000 records, it was

observed that the improvement in learning was limited — that is, adding more data had no significant impact on the convergence of the model. This experiment showcases the importance of choosing the right loss function for any specific application so that the model remains stable and reliable, regardless of the amount of data input.

The results present the basic need to find the right balance between detecting small errors and effectively dealing with outliers — especially when it comes to applications of nutritional information prediction. SmoothL1Loss acted as an aid for the model helping it generalize better across different food types, limiting overfitting, and maintaining high prediction accuracy. Because of the adaptability the function provides, an acceleration of the training process was observed, since the model was not negatively affected by unusual or extreme nutritional values. The overall experience confirms the importance of testing different loss functions, rather than just using the first available option, since there is never one correct option for every application. The final decision can have a decisive impact on the quality of training, the stability of the model, and — most importantly — whether it is able to perform well on real-world data.

Table 6 Comparison of Loss Functions for Model Optimization

Loss Function	Error Handling	Robustness	Convergence Stability
SmoothL1Loss	Balanced	High	Stable
MSELoss	Sensitive to outliers	Low	Unstable
HuberLoss	Moderate	Medium	Conditional Stability
L1Loss	Ignores small errors	High	Slower Convergence

3.5.2 Optimization Algorithm

One of the most important factors when training a model is the selection of the appropriate optimization function, as it directly affects both the performance and stability of the system. In this study, the AdamW optimizer was chosen due to its ability to effectively manage weight decay, helping to limit overfitting, while providing dynamic learning rate adjustment during training. The adaptive nature of AdamW allows the model to update its weights efficiently and

consistently, without abrupt changes, which leads to smoother and more stable training. Furthermore, it outperforms other optimizers, as it does not require extensive experimentation with hyperparameters, which is particularly useful when the model is complex.

Table 7 Comparison of Optimization Algorithms

Algorithm	Advantages	Disadvantages
AdamW	Adaptive learning rate	Higher memory consumption
SGD	Low memory usage	Slower convergence
RMSprop	Good stability	Sensitive to hyperparameter tuning

A critical hyperparameter that controls how much the model weights are updated with respect to the loss gradient during training, is the learning rate. In essence, it determines how fast the model tries to learn — and if it is not chosen carefully, the training can get out of control. We experimented with many values in the range of 1e-5 to 1e-3, and found that the best results were accomplished with 0.00012 (1.2e-4). This value turned out to be ideal as it allowed the model to learn quite quickly, without becoming unstable or exhibiting exploding/vanishing gradients. Table 8 demonstrates the impact of learning rates on model training. The experiments shown below differ only on the learning rate configuration, and it seems clear that moderate values such as 1.20E-04 perform better, achieving the lowest validation loss of 0.9537 and the highest R^2 score of 0.471. When this is considered alongside the performance of dataset sizes, it becomes evident that scaling up data volume does not linearly improve predictive accuracy. In fact, the model trained on a smaller, well-structured 10k sample achieved greater generalization than larger datasets, emphasizing that the strategic curation and preprocessing of data is more beneficial than raw data expansion. Having a curated dataset paired with optimal hyperparameter configuration leads to improved accuracy and reduces computational cost.

Table 8 Impact of Learning Rate on Model Performance

Learning Rate	Best Validation Loss	MAE	MSE	R^2 Score
1.00E-05	1.0344	1.2768	15.9826	0.4674
2.00E-05	0.9823	1.2219	15.2901	0.4566
1.00E-04	0.9843	1.2228	15.8906	0.4615
1.20E-04	0.9537	1.1777	14.9658	0.4710
5.00E-04	3.1718	3.4898	46.1799	-0.0643
1.00E-03	3.4033	3.7283	59.3563	-0.1987

Moving forward to the batch size, which determines how many training samples the model processes before updating its internal weights. In this study, 32, 64 and 128 were tested, and it was found that 64 was the most efficient, offering a good balance between training speed and final model quality.

An epoch, in model training, refers to one complete pass through the entire training dataset. In this study, experimentation started from 2 epochs, up to 12, providing the information that shorter epochs — [4, 6, 8] — reduced training time but failed to allow the model to converge appropriately, while epochs ofsize10 and 12, provided stability and prediction accuracy. Careful configuration of the hyperparameters is crucial, since any imbalance can cause unwanted results both on training efficiency and overfitting. Such configurations should be agreed based on experimentation on those hyperparameters and observation of the result, which will differ based on the specific application.

3.6 Performance Metrics and Evaluation

This section examines dataset size effects, model performance variations, and key evaluation metrics for predictive accuracy.

3.6.1 Dataset Size Influence

The size of a dataset greatly affects a model's performance — from its ability to generalize, to the accuracy of its predictions, to how quickly it processes the data. In this study, different dataset sizes were tested: 10,000, 80,000, 150,000, and up to 300,000 entries, to examine how the amount of data affects the model's performance.

The results were somewhat unexpected: smaller, more focused datasets, belonging to specific categories, almost always performed better than large, general ones. When a model is trained on concentrated, well-organized data, its predictions become more accurate and targeted. This shows that the quality and relevance of the data are much more important than the sheer quantity. Adding more data does not necessarily mean better results.

After about 80,000 records, a point of diminishing returns was clearly observed. Adding more data did not significantly improve the model; instead, it made training more unstable and less efficient. At the 300,000-record point, the validation loss increased relative to the training loss — a classic sign of overfitting. The model began to "remember" the training set rather than truly learning how to generalize to new data.

The trade-off between dataset size and model efficiency is crucial for real-world applications, where computational cost, inference speed, and data quality must be optimized. The findings reinforce the importance of categorization over sheer volume, as targeted datasets (such as 10k and 80k category-specific entries) achieve lower validation loss and higher predictive stability. Proper dataset curation, rather than indiscriminate data expansion, is essential for ensuring that the model captures meaningful patterns while maintaining generalization capability.

3.6.2 Evaluation Metrics

Evaluating the model's performance can be conducted through common metrics. In this study, we focused on the main three metrics: the Mean Absolute Error (MAE), the Mean Squared Error (MSE), and the Coefficient of Determination (R^2 Score).

MAE measures the average deviation between predicted and actual values. In everyday words, it shows, on average, how far the model's predictions are from the true values, but it carries the disadvantage of treating all errors the same, whether they are small or large. In contrast, the MSE severely penalizes large errors by squaring them. It can be useful since it highlights the values where the model prediction is far from reality, but in case of outliers it can give a lot of importance to the values. To evaluate the errors highlighted from the previous metrics, R^2 Score was used, explaining the model's variance in the dataset, as it offers a better overview of the overall performance rather than the error level.

Usage of all three metrics together enabled a much clearer and more comprehensive picture of the model's behavior, that is not coupled to the size of the dataset or the training conditions. This combination of metrics was one of the major factors that help us identify errors and enabled the continuous improvement of the model.

Table 9 Evaluation Metrics and Their Significance

Metric	**Purpose**	**Strengths**	**Limitations**
MAE	Measures absolute error	Easy to interpret	Ignores large error impact
MSE	Penalizes large errors	Highlights extreme deviations	Overly sensitive to outliers
R^2 Score	Measures variance explained	Provides model fit evaluation	Can be misleading for biased data

3.7 Machine Learning Classification and Feature Analysis

In this section, we analyze how DistilBERT embeddings can be used for nutrient prediction, focusing on regression problems rather than classification. In addition, we will also present different approaches to feature representation and the process of selecting the appropriate model for each case.

3.7.1 Feature Representation in Nutrient Prediction

The embeddings produced by DistilBERT represent a significant innovation in the way we transform ingredient lists into data that a model can "understand" and learn from. Rather than relying on older techniques — such as dictionaries or simple bag-of-words counting — DistilBERT creates deep, contextual embeddings that capture real-world associations between ingredients and their nutritional properties. A particularly useful feature of DistilBERT is its ability to detect similarities between ingredients, even when they are worded differently. By mapping ingredient lists into a high-dimensional space, the model begins to recognize subtle, semantic connections between ingredients and nutrients — something that no rule-based approach can achieve at this level.

Using pre-trained language models like DistilBERT is particularly useful in nutritional analysis, especially since ingredient names vary significantly across datasets. Transformer-based architectures are designed to generalize, allowing the model to make reliable nutrient predictions even when the data is not fully labeled. DistilBERT's true power comes when ingredient lists use different or ambiguous wording. Thanks to its learned associations, it can "fill in the blanks" and correctly predict nutrient values, even when the exact same terms do not appear. Furthermore, because it understands context, it can detect hidden patterns that simple word-based methods would miss.

Vectorization of features is essential for nutrient prediction with regression models. In cases where nutritional values are missing — which is very common — proper handling of these gaps is critical to the accuracy and stability of the model. Typically, imputation techniques are applied to replace missing values without altering the rest of the data set. Embedding-based methods also help by transferring the model's knowledge of similar ingredients to meaningfully "fill in" the gaps. When the data is properly structured and transformed, it is now possible to build

models that are scalable, stable, and capable of making meaningful predictions on real-world data.

3.7.2 Regression-Based Nutrient Prediction with DistilBERT

By applying DistilBERT, we transition from classification to regression is a significant change, as the model now predicts continuous numerical values for nutritional elements, rather than classifying foods into predefined categories. Classification limits predictions to labels such as “high”, “moderate”, or “low”, in contrast to regression, that allows for the prediction of integer values, a useful tool for food analysis, where diversity is high and differences do not easily fit into narrow boxes. Our choice of approach for predicting nutritional values provides great results, since foods contain complex and heterogeneous combinations of ingredients that require precise predictions. The model ability to learn and associate textual ingredient data with numerical nutritional values enhances its ability in making informed estimates even when it encounters completely new combinations. This means that the resulting predictions are detailed and realistic, reflecting the real differences observed in foods, and not just the average of a general category. With regression, the model can “see” and represent all intermediate states, without being limited to a few categories.

Another advantage is that retraining is not required for each new food type — regression models dynamically adapt to any list of ingredients. In this study, we focused on five key nutrients: total fat, protein, sodium, total sugars, and fiber, as they are key indicators for nutrition and food labeling. The ingredient embeddings from DistilBERT are mapped directly to the nutritional values through various regression models. MLP (Multi-Layer Perceptron) was employed, which has the ability to detect more complex and non-linear interactions between ingredients and nutritional values, that simple linear models cannot capture. The simple regression models, such as Linear and Ridge regression, mostly assume a linear relationship between inputs and outputs. Specifically, Ridge Regression can be useful for distributing weights in a balanced manner, to avoid sudden changes caused by very similar features. Linear Regression, even though it is common since it is a powerful statistical and machine learning

method, is too simple for ingredient list complexity. On the other hand, MLP introduces non-linearity, enhancing the model's ability to identify hidden patterns and dependencies in the data.

The transition from classification to regression requires more attention to feature design, but the benefits are significant: more flexibility, generalization to unknown products, and avoidance of categorization limitations. By combining appropriate feature design with contextual embeddings from DistilBERT, the model goes far beyond traditional rule-based methods. It is scalable, more interpretable, and better suited to real-world applications in the field of nutritional science. Furthermore, by cleaning and standardizing the features, the impact of small differences (e.g., changing an ingredient within the same category) is limited, while avoiding the loss of information caused by simply categorizing values. Finally, the application of normalization and regularization techniques enhances the stability of the model, regardless of the dataset size.

Chapter 4
Results and Findings

In this section the results of the study will be presented and discussed, with main focus on the network dynamics of the nutrient prediction model, the classification performance at various data scales, and the comparative evaluations of different training methods. Moreover, functional representations, feature transitions, and the effect of dataset size on the generalization ability of the model are analyzed.

4.1 Network Analysis

The term network analysis refers to the internal structure and behavior of a neural network during learning or inference. The relationships and interactions between different elements

within the network (such as features, layers, or representations), the stability, consistency, and transitions within the internal processes of the network are elements that evaluate the performance of a neural network. An important parameter to determine which components affect the accuracy of the generated predictions is the connection between the extracted features. Useful information regarding the consistency of the model in capturing the ingredient-nutrient relationships can be extracted from the transition probabilities between different representations, with changes indicating possible inconsistencies or biases in the learning process. Furthermore, the stability of the functional representations and their ability to remain reliable in various food labeling environments were assessed, highlighting the extent to which the model maintains its predictive capabilities when trained on general datasets versus categorized datasets.

4.1.2 Performance of Functional Representation in Food Labelling

Table 10 Configuration of model 3555bffe

Parameter	Value
Learning Rate	1,2e-4
Epochs	80,000
Batch Size	64
Loss Function	SmoothL1Loss
Sample Size	80.000
Categorization	Cheese

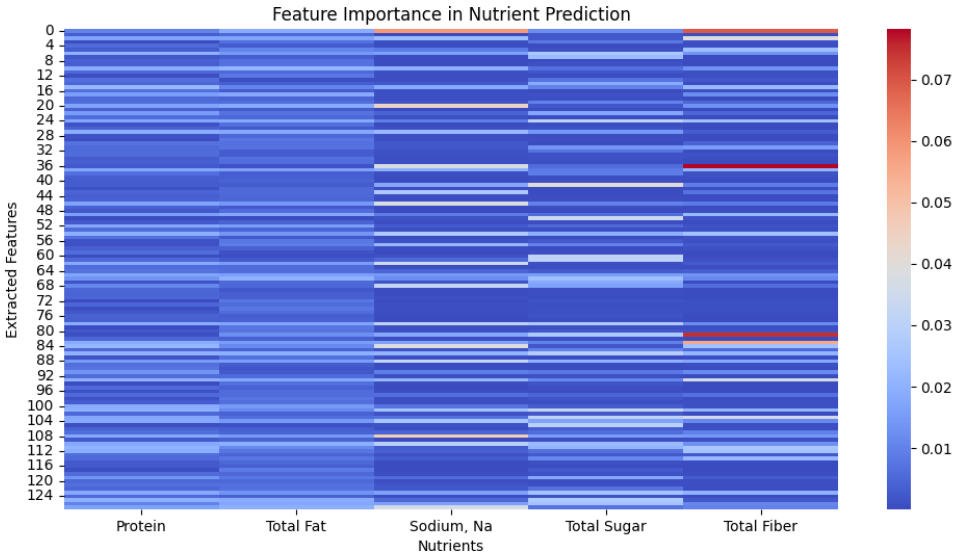

Figure 3: Feature Importance in Nutrient Prediction

Identifying which features really play a role in nutrient prediction is critical — especially when building a model to be both accurate and understandable. For this reason, a feature of importance heatmap was constructed to show at a glance how much each feature contributes to predicting the five key nutrients: Protein, Total Fat, Sodium (Na), Total Sugars, and Fiber. Given the configuration shown in Table 10, in Figure 3, features with high importance are shown in red, while less important features are shown in blue. This type of visualization makes it much easier to identify which features are “carrying” the weight of the predictions and how the predictive power is distributed. It also provides a picture of what the model has actually “learned,” as well as where there is room for improvement, through redesign or restructuring of features.

Analysis of the heatmap (Figure 3) indicates that specific features significantly contribute to nutrient prediction accuracy, particularly for macronutrients such as protein, sodium, and total fat. Therefore, features associated with frequent co-occurrences of ingredients possessing well-defined nutritional profiles are assigned greater importance by the model, underscoring the strength of ingredient-based representations. On the other hand, features that exhibit lower or variable importance indicate potential noise within the extracted set, pointing out the need for additional feature selection strategies or dimensionality reduction techniques to enhance model efficiency (Tiozon et al., 2023; Ahmed et al., 2024). One of the main reasons that the model

manages to perform so well is the use of contextual embeddings, which help to identify deeper relationships between ingredients and their nutritional profile (Pellegrini et al., 2021; Ispirova, 2022). Unlike traditional feature selection techniques, using fixed, rigid rules, embeddings generated from transformer models dynamically adapt to the structure of each data sample.

This gives the model the ability to generalize much better — even across completely different food categories (Rane et al., 2024). What's impressive about transformer models is that they can capture both the "big picture" and the small details, understanding not only the apparent relationship between ingredients, but also the underlying structure that connects them. This becomes even more apparent when you compare the model's performance on targeted, small datasets versus larger, more heterogeneous ones. Pretrained embeddings from models like BERT have already been shown to deeply understand semantic information of food ingredients, leading to better performance— whether it's food classification or nutritional value prediction (Devlin et al., 2019; Beltagy et al.,2020).

However, the variability in feature importance across different food categories points to a challenge coupled with using high-dimensional embeddings. Processed and branded food products offer a structured relation between ingredient embeddings and nutrient content, when in contrary, raw or minimally processed foods display greater variability, introducing noise and reducing model's robustness if not properly regularized (Akbari, 2023; Cao et al., 2024; Raffel et al., 2020). This variability points to that for a given and heterogeneous dataset, domain-specific adaptations need to be applied to avoid limitations while generating embeddings.

Overall, the feature importance heatmap gives a pretty clear picture of what works and what doesn't in terms of extracting useful information for predicting nutritional values. Contextual embedding is a big help — especially when the dataset is well-structured — as they allow the model to identify meaningful patterns. However, to fully exploit these methods on larger and messier datasets, the next step is to turn to hybrid approaches: combinations of domain knowledge and learning directly from the data. Such an approach can enhance both the accuracy of predictions and the interpretability of results — something that is extremely useful, both for scientific food analysis and for regulatory applications where transparency and reliability are essential. (Vaswani et al., 2017).

Based on the results of the feature importance analysis and the model's behavior, the right feature selection can make a significant difference in the accuracy of nutrient prediction. By focusing only on the most important features, complexity is reduced, the potential for overfitting is limited, and the overall speed of the system is improved. The feature importance heatmap (Figure 3) shows this in a visual way, since it reveals which features contribute significantly to the predictions and which ones are there without providing any real value. When you look at these results, it becomes clear that some features have a much greater impact than others. So, you can safely remove the less useful ones without negatively affecting the model's performance. This type of analysis is particularly useful: it helps you identify which variables are worth keeping for better classification, and which ones only add noise or affect the stability of the model.

Feature selection plays an important role in improving the accuracy of classification models — especially in applications like food classification based on nutritional value, where datasets can become huge and noisy. When you have too many features, it's very easy for a model to overfit or simply become slow and dysfunctional. The literature on the importance of features clearly shows how difficult it is to strike the right balance. You want your model to be powerful enough to recognize meaningful patterns, but also simple and transparent enough so that you can understand what it's actually doing (Folorunso et al., 2023; Ibrahim et al., 2024). In our study, the heatmap reveals that a small subset of features carries disproportionate importance, confirming that eliminating less relevant attributes could enhance the model's ability to generalize across different datasets (Chandrashekar & Sahin, 2014). However, indiscriminate removal of features may lead to a loss of critical information, highlighting the necessity of careful validation techniques in feature selection processes.

The results clearly show that there is a delicate balance between selecting the right features and maintaining the simplicity of the model (Neloy & Turgeon, 2024). Other studies have shown that adding too many features — especially in high-dimensional data — only creates redundancy, which increases variance, burdens training, and leads to overfitting (Bolón-Canedo et al., 2015). In this study, the use of smart feature selection techniques, such as recursive feature elimination and mutual information ranking, had a catalytic effect: the models generalized better, without losing classification accuracy (Xue et al., 2015; Theng & Bhoyar, 2024). These improved

performance results demonstrate that when features are carefully selected, the signal becomes clearer, noise is reduced, and the model can classify nutrients with greater confidence (Garg & Dwivedi, 2024).

What is particularly interesting is that feature selection not only affects accuracy; it also plays a crucial role in the stability and robustness of the model when applied to different food categories (Zerouali et al., 2024). There are now several studies showing that models trained on smaller but more meaningful feature sets are better at responding to new, unknown data — something that is extremely important in the field of nutritional science (Ropodi et al., 2016; Jablonka et al., 2020). Our results confirm this trend: when limiting the data sample to the most appropriate and "clean" set of features, the model not only becomes better at classifying nutrients, but it also becomes less vulnerable to random or meaningless patterns (Ming et al., 2022; Ye et al., 2024). That's important in nutritional value categorization; interactions between features can easily blur the truly important signals — and lead the model to focus on secondary or misleading associations.

4.1.3 Feature Transition Probabilities and Variability Across Datasets

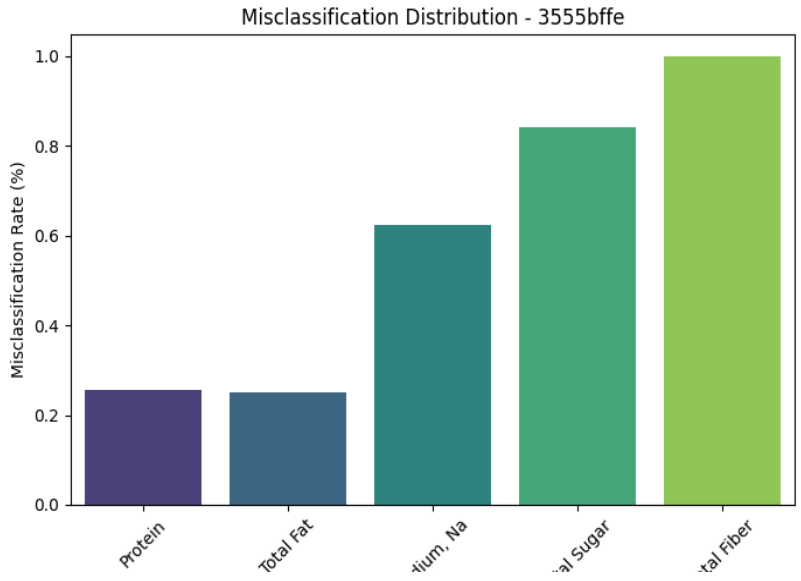

Figure 4: Misclassification Distribution Across Nutrients

Figure 4 illustrates the misclassification rates for different nutrients in the predictive model, with the configuration shown in Table 10. Higher misclassification rates for sodium, total sugar, and total fiber suggest challenges in distinguishing these nutrients due to feature overlap and dataset variability.

Observing how features change depending on the type of dataset — for example, by comparing category-specific sets with large, generalized ones — provides very useful insights into how the model is fitting. What has been found is that some ingredient-based features have a much greater impact on nutrient value predictions when using targeted datasets. However, their importance decreases significantly when the model is trained on general, more heterogeneous sets (Rastegar et al., 2023; Singhal et al., 2023). In other words, the way features are grouped directly depends on the structure of the dataset. In more targeted data, the associations between ingredient embeddings and nutrient values are much stronger and more stable (Barbiero et al., 2020; Naravane, 2024). In contrast, when a large range of variation is introduced, as occurs in general datasets, the diversity creates fluctuations in the importance of features, which reduces the stability of predictions (Van Giffen et al., 2022). This is in full agreement with previous studies that support that models perform better on nutritional tasks when trained on domain-specific datasets, rather than general, mixed sets (Mehrabi et al., 2021; McElhinney, 2024).

The misclassification analysis reveals substantial discrepancies in prediction accuracy among different nutrients, as illustrated in Figure 4. Sodium (Na), total sugar, and total fiber exhibit the highest misclassification rates, indicating that these nutrients are more prone to misrepresentation due to feature ambiguity and data sparsity. The transition probabilities suggest that similar ingredients across food categories may contribute to errors, particularly when nutrient profiles overlap (Choudhury, 2025). These inconsistencies highlight the need for refined feature selection techniques to mitigate classification errors and improve generalization (Budach et al., 2022).

Looking deeper into how ingredients cluster together, it appears that the nutrients that the model struggle with the most — such as sodium — often share similar ingredient profiles. Sodium in our study is a perfect example, as it appears in a wide variety of processed foods, but ingredient lists are not always detailed enough since such nutrients "hide" in almost every ingredient listed

and the quantity of it varies, making it difficult to balance flexibility with precision in identifying the most important features for each nutrient value (Mavrogiorgos et al., 2024). The accuracy of predictions can be optimized by leveraging techniques such as hierarchical clustering and adaptive feature weighting, as proposed by relevant studies (Tao et al., 2020). The fact that feature transition probabilities can change so dramatically from one set to another highlights just how complex nutrient prediction actually is, and the need for more adaptive models that can understand these variations (Delfani et al., 2024). We need to point out that by improvement of feature selection, and application of intelligent interventions per dataset, we can increase the reliability of classification and reduce prediction errors. For future work, as supported by relevant studies, it would be of particular interest to use reinforcement learning to automatically adapt the importance of features depending on the dataset, in such way, prediction algorithms could more effectively handle ambiguous, complex, and "noisy" data — a phenomenon common in large datasets — thereby reducing incorrect predictions (Hassler et al., 2019).

4.1.4 Temporal Convergence of Model Training

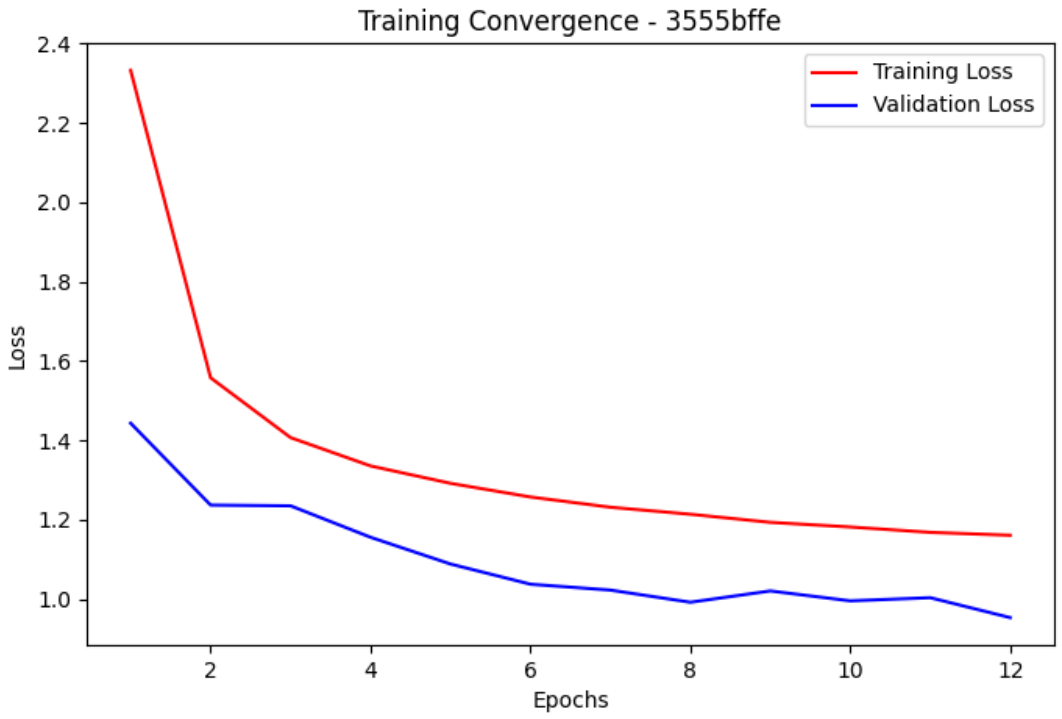

Figure 5: Training Convergence Across Epochs

Given the configuration of the model as shown in Table 12, we observe the behavior of the loss function during training. It becomes clear that the value of the training loss decreases sharply within the first few epochs and then stabilizes, which is a clear indication that the model learns effectively from the initial stages, as the parameters are adjusted correctly and the convergence process starts smoothly (Arjovsky, 2020; Cheng et al., 2024; Siddique et al., 2024). The picture is similar for the verification loss, which initially follows a downward trend, however, in the later stages of training it often fluctuates, which may indicate the beginning of overfitting of the model to the training data, especially when the values do not continue to improve at a constant rate (Roelofs et al., 2019). Monitoring these trends, both for training and validation, is particularly useful not only for evaluating the learning progress of the model but also for determining the optimal time point to stop training, thus ensuring high accuracy without unnecessary consumption of computational resources or unnecessary continuation of the process (Poldrack et al., 2020).

Looking at the accompanying diagram (Figure 5), it is clear how the loss changes with each epoch, displaying a sharp drop at the beginning which means that the model is learning quickly, while the leveling off of the training loss indicates that it is approaching convergence. Conversely, in case of fluctuations of the validation loss towards the end act as a warning that regularization may need to be strengthened, to avoid over-specification of the training data, which is not a fact for our model, displaying convergence once again. One of the main problems in predicting nutritional values with deep learning is maintaining the stability of the model, especially when the size of the dataset changes. If the validation loss continues to show strong instability in the last epochs, it may be due to low-quality features that affect generalization (Chattopadhyay et al., 2020). In parallel, we need to emphasize the role hyperparameters, as batch size and learning rate play, as they control the sensitivity of the model to weight changes (Freiesleben & Grote, 2023), with our settings of batch size at 64 and learning rate at 1,2e-4. Those hyperparameters need to be fine-tuned to keep the loss low without sacrificing the generalization ability of the model (Mienye & Swart, 2024).

The way the loss evolves per epoch provides valuable insights about how effectively the model is learning at each stage of training. Applying techniques such as early stopping can help

terminate training early, especially when the validation loss stops showing substantial improvement (Cui et al., 2018). Such approaches are useful for determining with greater certainty when the model has truly converged. Because even if there is sufficient computing power to continue training for multiple epochs, this does not necessarily mean that performance will improve. Furthermore, when the data or features are particularly complex, incorporating a dynamic learning rate adjustment mechanism — widely known as schedulers — can significantly enhance the stability of training (Montesinos López et al., 2022). With all these tools — from careful parameter tuning to proper monitoring of loss graphs — it is possible to maintain a model that is accurate, efficient, and stable, without introducing unnecessary complexity (Baer, 2015).

4.3 Comparative Evaluation of Training Approaches

The results of this section clearly show how important the choices made during training are, since they affect everything from how quickly the model converges to how well it generalizes and how it responds to noisy or unpredictable data.

First, we looked at how batch size affects performance. And we highlight a trade-off. Smaller batches help the modelgeneralize better, avoiding overfitting, while verylarge batches cause it to stabilize prematurely — making it less flexiblewhen encountering new data (Santos & Papa, 2022). The validation loss curves confirm this: large batches bring fast convergence in the first epochs, but lose in generalization as training progresses. So we conclude to this dilemma — speed or robustness when developing a model for the real environment (Wang et al., 2022; Suddul & Seguin, 2023).

Moving on to the loss functions, SmoothL1Loss,MSELoss, L1Loss, and HuberLoss were tested. Here too, the choice makes a difference. SmoothL1Loss seems to strike the right balance, since it keeps the model stable, robust to outliers, and avoids the excessive deviations caused by MSELoss, which often overestimates outliers (Ciampiconi et al., 2023; Ikram & Aslam, 2024; Yang et al., 2024).

Of course, the structure of the dataset is equally crucial. Specialized, category-specific datasets consistently perform better than large, generalized ones. Even beyond 80,000 records, the

improvement is marginal — and often performance deteriorates, due to additional noise and redundancy (Bottou et al., 2018; Yanet al.,2024). The results in R^2 and MAE clearly show that well-structured ingredient data yield much more reliable predictions, compared to the “more is better” approach (Zhang et al., 2024).

All of this together highlights how important it is to properly tune hyperparameters and curate your dataset if you want a model that really performs welland predicts accurately. In the future, it would be of particular interest to examine hybrid trainingstrategies that automatically adjust parameters based on the characteristics of the dataset — so that the model remains stable but flexible, depending on the food category it is applied to (Naumenko et al., 2024; Olufemi-Phillips et al., 2024).

4.3.1 Influence of Learning Rate on Convergence

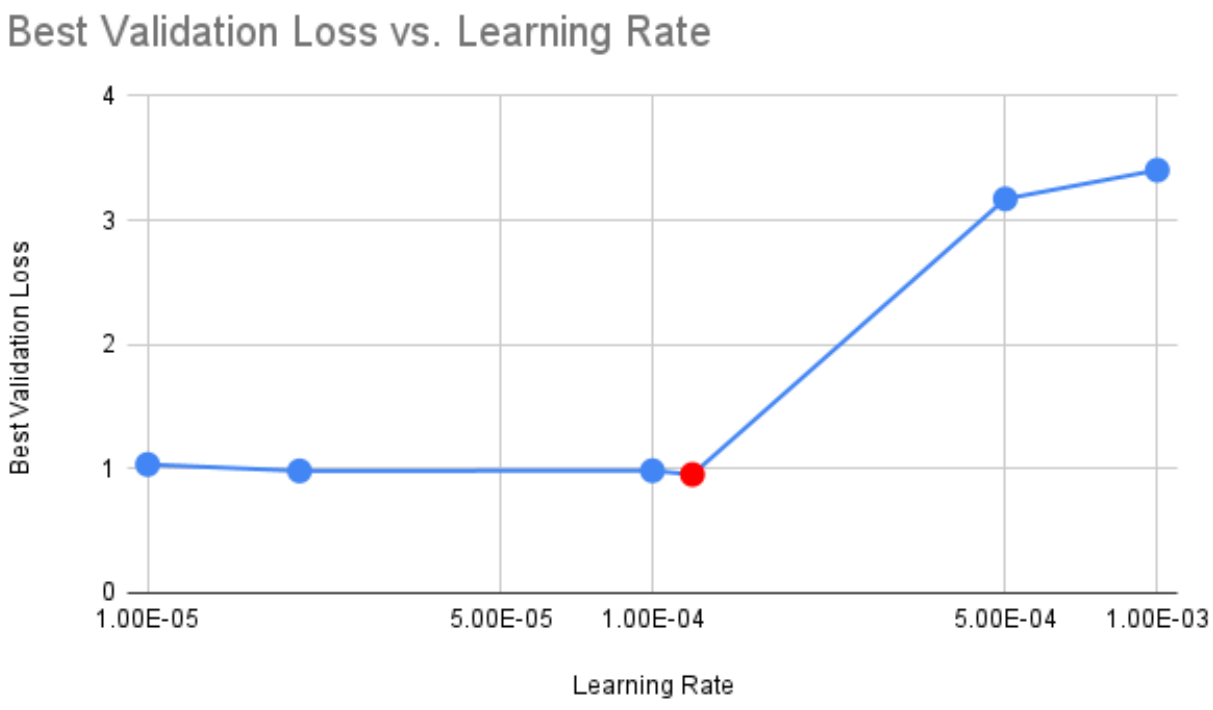

Figure 6: Best Validation Loss vs. Learning Rate

Choosing the right learning rate is critical when training neural networks — it can determine not only the stability of the process, but also how well the model generalizes. To find the optimal

value, tests were performed with learning rates ranging from 1e-5 to 1e-3, and the results are summarized in Figure 6 (“Best Validation Loss vs. Learning Rate”) and Table 9, where some recurring patterns are clearly visible.

As shown in Figure 6, lower learning rates (1e-5 to 1e-4) kept the validation loss at a stable and controllable level, while performance improved further at a value of 1.2e-4. In contrast, when the learning rate was increased to 5e-4 or more, the validation loss worsened significantly, indicating that the training became unstable and the model was likely to exceed the optimal convergence point. The optimal value (1.2e-4) is marked with a red dot in the graph — as it achieved the lowest validation loss (0.9537), achieving the ideal balance between fast learning and steady progress.

Table 9 confirms this choice numerically: at the value of 1.2e-4, the Mean Absolute Error (MAE) was 1.1777, the Mean Square Deviation (MSE) was 14.9658, and the R^2 index reached 0.471. At all levels — accuracy and efficiency —this value outperformed the others.

For this reason, 1.2e-4 was used in all subsequent trainings. It was consistently the most efficient choice, without leading to underfitting or instability, and had a direct, positive effect on the overall performance of the model.

Table 11: Comparison of Validation Metrics Across Learning Rates

Learning Rate	Best Validation Loss	MAE	MSE	R^2 Score
1.00E-05	1.0344	1.2768	15.9826	0.4674
2.00E-05	0.9823	1.2219	15.2901	0.4566
1.00E-04	0.9843	1.2228	15.8906	0.4615
1.20E-04	**0.9537**	**1.1777**	**14.9658**	**0.4710**
5.00E-04	3.1718	3.4898	46.1799	-0.0643
1.00E-03	3.4033	3.7283	59.3563	-0.1987

4.3.1 Influence of Batch Size and Epochs on Convergence

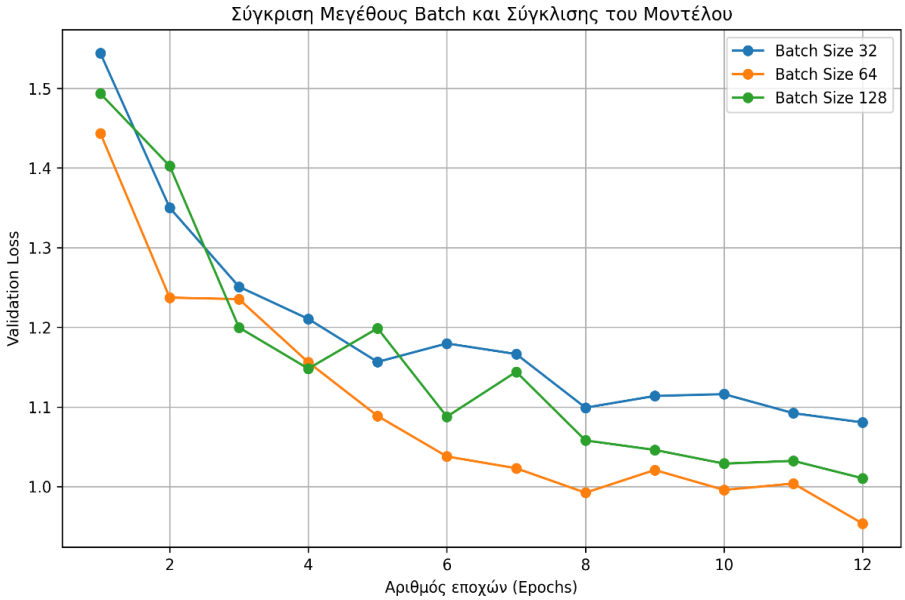

Figure 7: Comparison of Batch Size and Convergence

As we clearly see in Figure 7, the batch size through a number of epochs affects the model's convergence. A clear pattern is observed: smaller batch sizes, such as 32, may initially generate more noise, but in the second half of training they lead to more stable and reliable predictions. On the other hand, larger batches, such as 128, accelerate the decline of loss in the first epochs, but often struggle to generalize effectively to new, unknown data (Smith et al., 2021).

As it has been observed that smaller batches help the model escape from sharp local minima, improving generalization (Keskar et al., 2017; Do et al., 2024), with the trade-off is that very small batches can slow down training and lead to unstable gradients due to excessive stochasticity.

Another critical factor is matching the batch size to the scale of the dataset. As the dataset grows, it makes sense to increase the batch size to maintain efficiency and avoid instability in weight updates (Goodfellow et al., 2016; Menghani, 2023). However, after a certain point, further increasing the batch size does not provide any significant benefits and may lead to superficial local minima and weaker generalization (Hoffer et al., 2017; Geiping et al., 2021).In

the context of this study, tests on datasets of different sizes showed that batch size 64 was the optimal balance point, while larger values either did not improve performance or worsened validation loss.

At the same time, the application of early stopping seems to be particularly useful — providing to the model this automated ability to stop training at the right time saves time and improves its generalization ability. The loss curves clearly showed that after a certain number of epochs, performance stabilizes or even deteriorates, due to overfitting (Van Leeuwen & Nutzel, 2024; Yu et al., 2025). We have to keep in mind that the appropriate stopping point varies depending on the size of the dataset — smaller datasets require shorter training times, while larger datasets can withstand more epochs before overfitting occurs (Dinkel et al., 2021).

The findings confirm the need to adapt the batch size according to the characteristics of each dataset, in order to achieve meaningful and efficient parameter updates, without compromising the generalization ability, since with the correct choice of batch size and the timely application of early stopping largely, we could better determine the stability and accuracy of the final model. Interest for future work, is highlighted in the field of adaptive batch sizing, where the batch size is dynamically adjusted during the training phase of the model (Goyal et al., 2017). Such an architecture, could drive significant improvements in efficiency, in fields similar to nutritional prediction, where the data are characterized by high heterogeneity and complexity (Ma et al., 2022; Armand et al., 2024).

4.3.2 Loss Function Comparisons and Model Stability

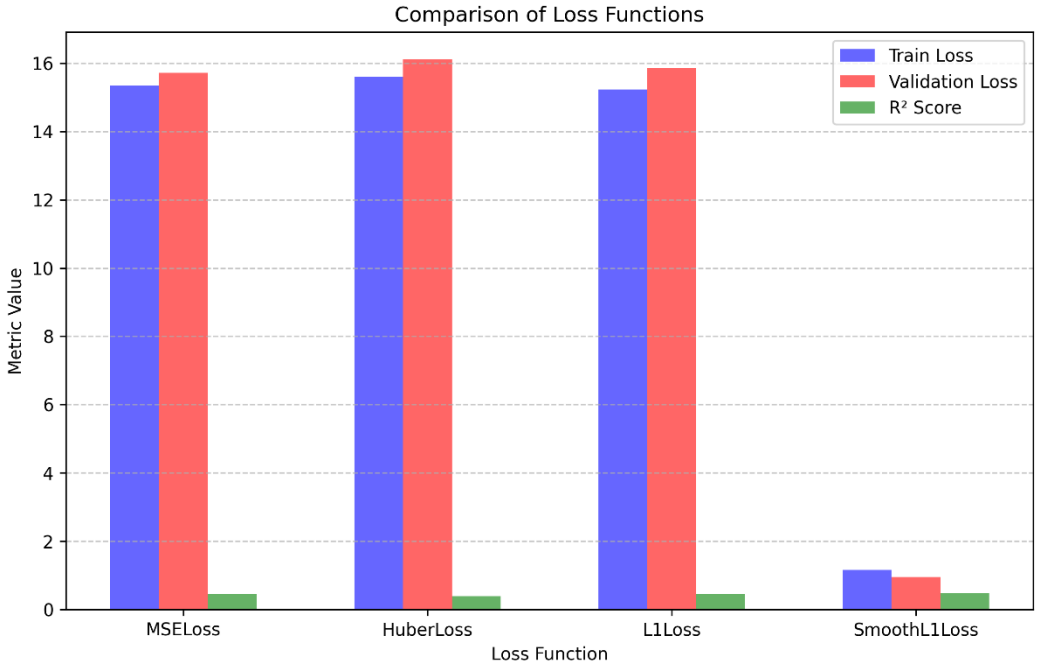

Figure 8: SmoothL1Loss achieves the lowest validation loss

Given the large amount of data we are required to manage in the field of nutrition, it was clear that choosing the right loss function would make a significant difference in the generalizability of the model. In our comparative study, four widely used functions — SmoothL1Loss, MSELoss, L1Loss, and HuberLoss — were examined indepth, with the aim of determining how they deal with outliers and how they affect the overall performance of the model. SmoothL1Loss clearly emerged as the most reliable choice, managing to balance stability and accuracy where the others showed inconsistency, without the need for further parameter tuning as required in the case of HuberLoss. More common choices, such as MSELoss and L1Loss, proved to be less stable — either over-penalizing large deviations or ignoring small but significant differences. This led to unstable predictions and lowoverall generalization ability of the model. In contrast, SmoothL1Loss exhibited adaptive behavior, combining the advantages of the others without getting trapped in their disadvantages. The result was less overfitting, better generalization, and

— most importantly — predictions that remained reliable even in complex, “unstructured” real-world data. For models that requirestability but high resolution, SmoothL1Loss appears to be a powerful model for designing predictive systems (Tsoy et al., 2024; Wang et al., 2024).

Studying the convergence behavior of these loss functions per epoch shows that models trained with SmoothL1Loss stabilize faster compared to those using MSELoss, indicating that the adaptive nature of the former absorbs extreme slope values that could otherwise destabilize learning. When comparing loss functions, we notice that HuberLoss stands out for its stability, especially when applied to datasets that include a mixture of structured and unstructured variables. In such data— where homogeneity is lacking and outliers are common — HuberLoss manages to adapt to fluctuations without becoming destabilized (Meng et al., 2024; Kabir et al., 2025). However, this adaptability does not come without a cost. Unlike more “off-the-shelf” functions, HuberLoss requires careful tuning of the parameter δ (delta), which determines when the loss will transition from quadratic to linear behavior. This step, while providing accuracy, can slow down the initial model tuning phase, as it requires additional testing and fine-tuning — especially when the dataset is large or heterogeneous. Figure 8 confirms that the validation loss is much smaller with SmoothL1Loss than with MSELoss, or the rest, supporting the hypothesis that over-penalizing large errors leads to instability. This is in line with previous studies, which show that although MSELoss is effective on well-formed data distributions, it amplifies errors when unpredictable deviations appear.

When it comes to a model’s ability to generalize, the choice of loss function plays a crucial role — not only in how quickly the model learns, but also in how reliably it performs over time. When looking at R^2 values we can see that SmoothL1Loss scores a bit higher, but consistently, and proves to be the best choice on both general and filtered datasets. In contrast, MSELoss and L1Loss, continuously managed to generate huge validation and training losses, regardless of the dataset fed to the model, making them a bad choice for our study. It is important to note that HuberLoss, was designed primarily to limit the effect of outliers and exhibits higher overall predictive ability than SmoothL1Loss, making it preferable when accuracy is a priority (Mathew et al., 2024; Yaqoob & Muntean, 2024), but the additional step of adjusting the δ parameter to avoid performing similarly to the other three, proves to be very time-consuming and not worthy

in our case, given the huge gap on the metrics SmoothL1Loss manages to achieve from the very start.

As a conclusion from this comparison, we understand that the optimization criteria must be chosen with a clear awareness of both the specificities of the dataset and the purpose of the application. We cannot exclude MSELoss and L1Loss, since they remain reliable benchmarks in fully controlled environments, their use in noisy environments requires special care (Kosma, 2023; Lei et al., 2025). In practical applications such as ours, SmoothL1Loss stands out as the most stable and flexible choice, since it has the ability to handle both small and large deviations while keeping strong generalization ability. HuberLoss, while useful in specific environments, fails to compete with SmoothL1Loss as a general-purpose solution, given the time needed to fine-tune it to stand out from the rest.

4.3.3 Generalization Ability of Category-Specific vs. Large Datasets

Table 12 Configuration of model while testing sample size and categorization effects

Parameter	Value
Learning Rate	1,2e-4
Epochs	80,000
Batch Size	64
Loss Function	SmoothL1Loss

The experiments that highlight the importance of sample size and focused datasets are presented in Tables 13 and 14 with their evaluation metrics. It is worth mentioning that all the experiments are conducted with the same configuration, shown at Table 12, and they only differ on sample size and dataset categorization, making it clear that when trained on smaller, category-specific datasets (e.g., 10,000 and 80,000 entries), achieved significantly lower validation loss and Mean Absolute Error (MAE), demonstrating benefits in terms of thematic organization of the data. As the sample size increases to 150,000 and 300,000 entries, only small improvements in R^2 values are observed, while the decreases in MAE and MSE are minimal. This suggests that there is a capacity point at which adding more data enhances memorizing over predictive ability.

Such patterns, where the train loss value is lower or equal to the validation loss, suggest potential overfitting or increased sensitivity to noise in larger, more heterogeneous datasets, consistent with concerns about reduced generalization in high-variance environments (Budach et al., 2022). The consistency of improved results in smaller, curated datasets reinforces the idea that well-defined feature distributions contribute to training stability and model coherence (Chatterjee & Zielinski, 2022; Freiesleben & Grote, 2023). In nutrient prediction tools, where ingredient complexity and inconsistent labeling pose challenges to model training, limiting the data size with structured inputs proves more effective. This reinforces the broader argument that strategic curation of datasets, over simple expansion of the sample size, is the way to create reliable and interpretable machine learning models (Choudhury, 2025).

Table 10 Effect of Sample Size on Validation Loss and Predictive Metrics using non-categorized products

Experiment ID	Sample Size	Best Validation Loss	Final Train Loss	MAE	MSE	R^2 Score
2bb81b72	10,000	2.3097	1.7938	2.6781	44.8373	0.5001
bf14f615	80,000	1.6490	1.4564	1.9671	31.5877	0.7304
910e736f	150,000	1.5102	1.3832	1.7978	28.1527	0.7607
7cd5bb06	300,000	1.3776	1.3787	1.6988	27.6342	0.7797

Table 11 Effect of Sample Size on Validation Loss and Predictive Metrics using categorized products

Experiment ID	Sample Size	Best Validation Loss	Final Train Loss	MAE	MSE	R^2 Score
144c11cc	10,000	1.3051	1.4712	1.6389	18.0403	0.358
3555bffe	80,000	0.9537	1.1612	1.1777	14.9658	0.471

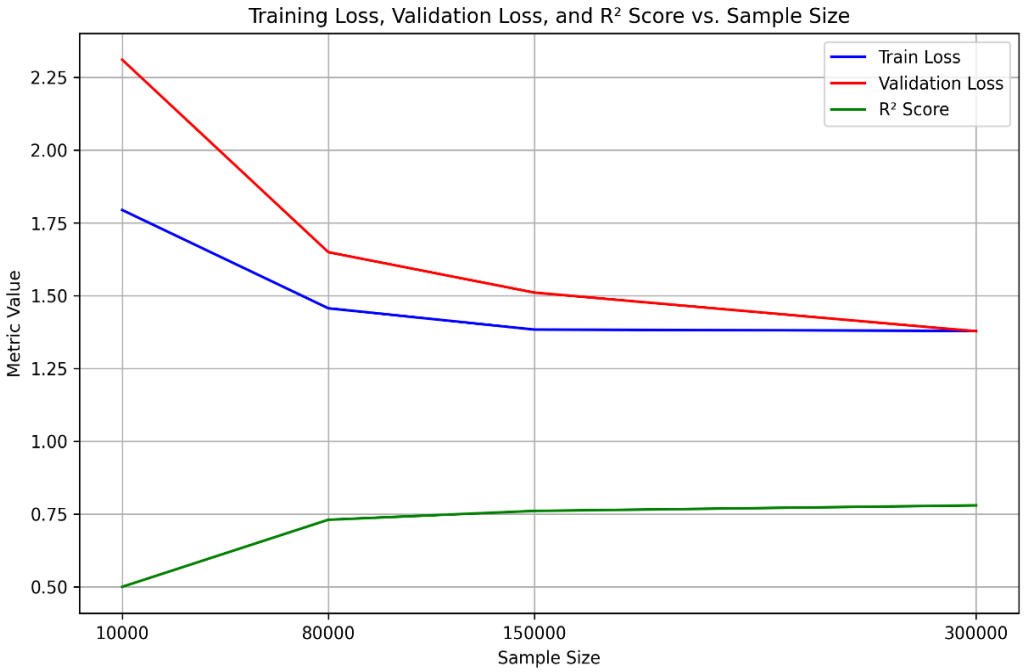

Figure 9: Impact of Sample Size on Loss and R^2 Score

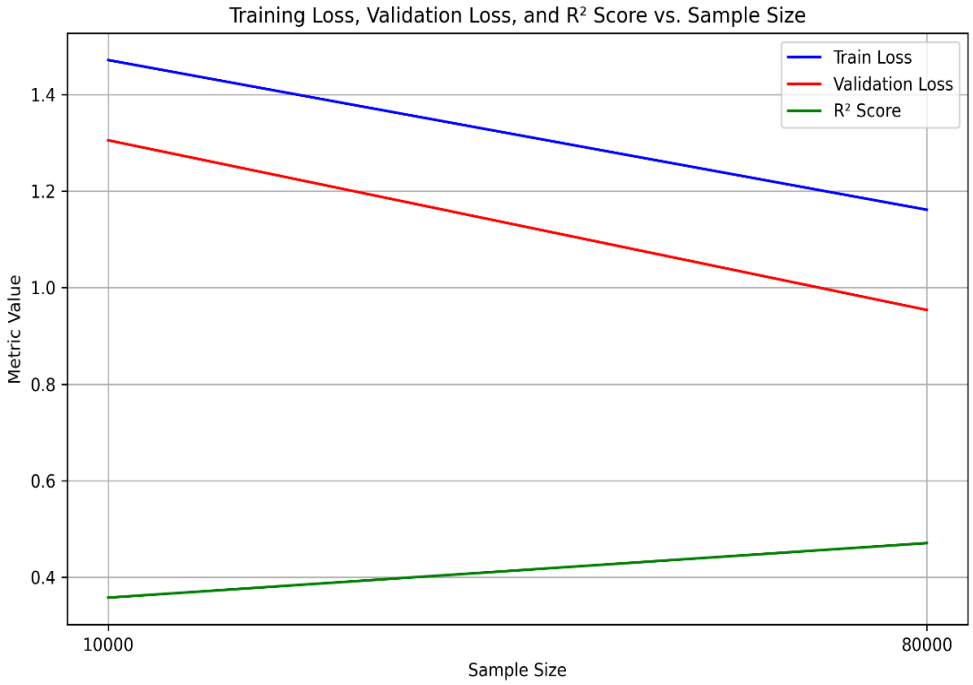

Figure 10: Impact of Sample Size and Categorization on Loss and R^2 Score

The analysis shows that smaller, carefully selected datasets lead to better generalization compared to larger, unfiltered datasets — especially when the structure of the ingredients follows a specific food category classification. The results, shown in Figures 9 and 10, are in line with previous research showing that well-curated training sets enhance the model's ability to recognize meaningful patterns while limiting overfitting, especially when the noise in the dataset is controlled (Ciampiconi et al., 2023). However, smaller dataset have their own limitations, since filtering a larger dataset to create a smaller and specific one, we limit the model's overall knowledge, leading it to have reduced accuracy when it comes to unknown data that might encounter in real-world scenarios.

On the other hand, large datasets have the benefit of broader coverage enablingthem to be used in more real-world scenarios, with the downside that, for the same reason, they also introduce noisier predictions which are not that accurate (Wang et al., 2022). This issue is particularly critical in the field of machine learning, where noise and class imbalance can

significantly affect predictive accuracy (Santos & Papa, 2022). Categorization of the data sample applied with caution, since in nutrition there are cases where the data category may consider exhibiting similar data but in reality the composition varies significantly, lowering this way the confidence level of the model while predicting. Numerous studies confirm that both the structure of the dataset and the selection of features help in reducing noise and stabilizing model performance (Thakkar & Lohiya, 2022; Ciampiconi et al., 2023; Theng & Bhoyar, 2024).

All the above findings support the hypothesis that structured datasets contribute to the improvement of the generalization ability of a model. Models trained on refined datasets show better alignment with real-world nutritional values, highlighting the importance of balancing size and structure when optimizing predictive accuracy without sacrificing usability in real-world settings. We managed to clearly demonstrate that models trained on a wide range of foods with high category diversity exhibit reduced predictive consistency when, on the other hand, data that are carefully filtered contain well-structured ingredient lists, and their performance metrics display increased accuracy, indicating that structured datasets facilitate the adaptation of the model to trends related to the content itself.

4.4. Validation Results and Comparative Metrics

Reminding the results, we showed that SmoothL1Loss outperforms other options, showing lower loss during both training and validation, which indicates a better fit of the model to the data (Miraftabzadeh et al., 2018; Weng, 2020). In contrast, traditional approaches such as MSELoss and HuberLoss show higher loss values, which indicates increased sensitivity to outliers or noise in the data (Jadon et al., 2024). Particularly important is that the R^2 score is lower in cases where MSELoss is applied, which indicates that the model fails to accurately capture the fluctuations of the input data, negatively affecting its performance (Ikram & Aslam, 2024). In our study, the choice of SmoothL1Loss achieved balance between noise.

As discussed, when comparing the loss functions, the choice of the appropriate function can significantly affect the reliability of the predictions, where in our study, the use of SmoothL1Loss seems to offer a balance between handling noise and maintaining sensitivity to real data changes, which made it more suitable for applications where outliers should not disproportionately affect

the behavior of the model (Al-Huthaifi et al., 2024; Ishwarya & Kothandaraman, 2024). At the same time, the HuberLoss and L1Loss functions show better performance compared to MSELoss, but they do not reach the level of stability offered by SmoothL1Loss (Ciampiconi et al., 2023).

4.4.1 Model Evaluation Using MAE, MSE, and R^2 Score

Evaluating model performance goes well beyond checking if the numbers line up. In this study, three core metrics — Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 (the coefficient of determination) — each shedding light on a different aspect of predictive accuracy. Higher MAE and MSE scores signaled that the model was missing the mark, struggling to match predictions to reality — especially when underlying data patterns were elusive. Lower MAE and MSE scores indicated a good result, usually showing up when models were carefully tuned. R^2 scores brought another layer of insight: the higher the value, the more variance the model could explain. What really stood out was that models trained on well-structured, domain-specific data almost always posted higher R^2 scores, making a strong case for the value of targeted preprocessing (Ciampiconi et al., 2023; Wang et al., 2022). A closer look at MAE and MSE under different loss functions showed just how much error penalization can impact results. Models trained with SmoothL1Loss, for example, tended to produce lower MAE than those using traditional MSELoss, thanks to their ability to brush off the influence of outliers (Kaviani et al., 2024; Yang et al., 2024). MSELoss, by contrast, was quick to spike in the presence of noise or extreme values — useful for catching big misses, but sometimes less helpful when it came to practical, real-world predictions (Dessain, 2022; Hosamo & Mazzetto, 2024). The research sets clear that picking theright loss function, and pairing it to your data, is essential for makingmodels both stable and reliable (Santos & Papa, 2022).

Variation in R^2 across training runs also highlighted the huge impact of dataset quality and model architecture (Ayman et al., 2024). Curated, category-specific datasets consistently beat out massive, unfilteredones — proving that smart feature engineering really pays off (Koido et al., 2023; Mostafa, 2024). High R^2 meant the model could generalize well without overfitting, which is critical for real-world prediction tasks (Ma et al., 2022). Still, even with these improvements, some tricky test cases stuck out, hinting that more work on feature weighting or regularization could boost overall stability (Ciampiconi et al., 2023). And while bigger datasets

sometimes nudged up validation scores, the returns were often small — suggesting that simply scalingup isn't always the answer (Kosma, 2023).

The side-by-side comparison of MAE, MSE, and R^2 shows that model evaluation is anything but one-dimensional, with each metric offering its own perspective, and picking the right mix depends on the job at hand (Menghani, 2023). R^2 is great for a quick sense of model fit, but it needs to be weighed alongside error-based metrics for a full picture. We conclude that, the balance between loss functions, data curation, and evaluation metrics forms the backbone of any effort to optimize predictive models for real-world use (Goyal et al., 2017; Davis et al., 2024). Going forward, the evidence suggests that tuning loss functions to the data will be key to building models that generalize better and aren't thrown off by oddities or noise (Faber et al., 2024).

4.4.2 Model Evaluation using Bland-Altman Plots

In this section of the study, we will examine the strengths and weaknesses of the model when it tries to predict each specific nutrient. The results are produced from the output of the validation set, of the data sample, with configuration of Table 12. Analysis of the Bland-Altman plots provided useful insights on the generalizability and accuracy of the model, for each specific feature, indicating the effect of retaining or removing features in shaping the model's performance. A Bland-Altman plot, visualizes the degree of agreement between the predicted and actual nutrient value, providing a detailed picture of both the consistency of the model and the biases or deviations introduced.

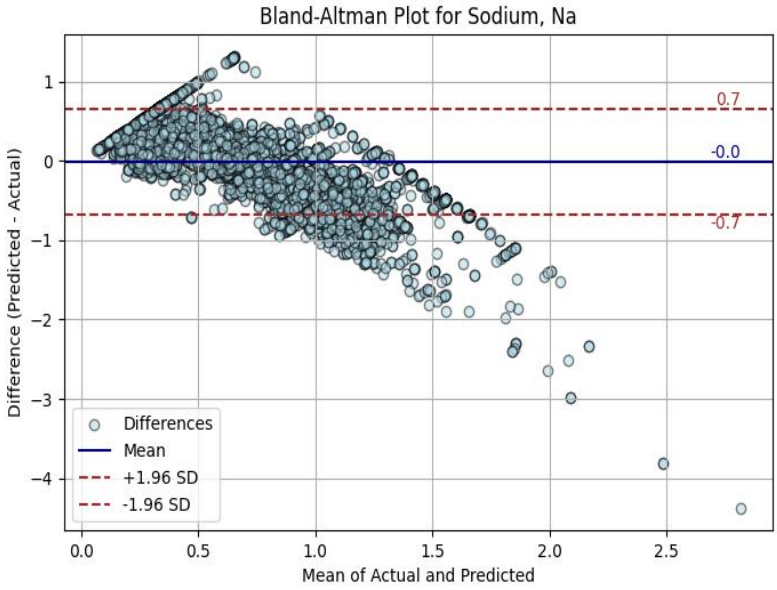

Figure 11: Agreement of Sodium Prediction vs Actual

Figure 11 demonstrates the presence of systematic bias as well as increasing dispersion at higher sodium values. The wide distribution of points showcases some inconsistencies in the predictions, with larger deviations at the extreme values, highlighting tendencies of overestimation or underestimation by the model, which may require further calibration.

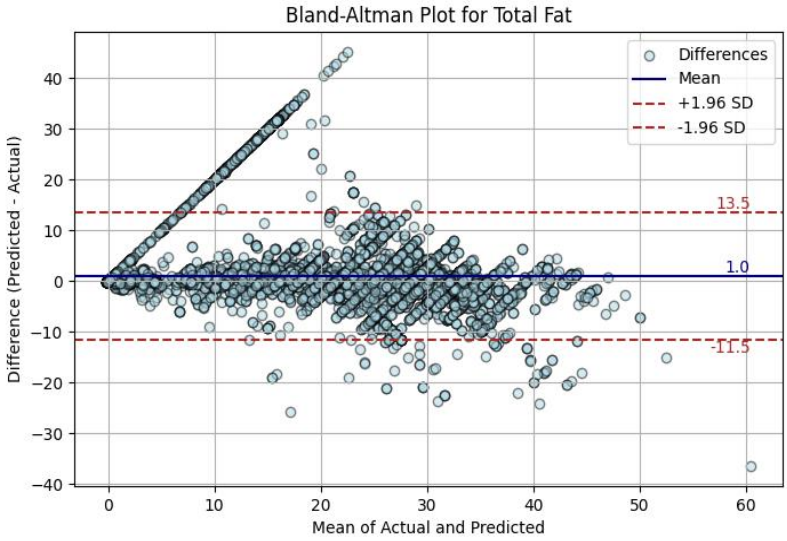

Figure 12: Agreement of Total Fat Prediction vs Actual

In Figure 12 we can predict that the majority of the predictions fall within the limit of agreement. It is clear that, we have greater prediction errors at higher fat levels, but with the mean difference between predicted and actual values close to zero, indicating little to no systematic error. It also displays a few outliers but with no clear tend to increasing or decreasing error.

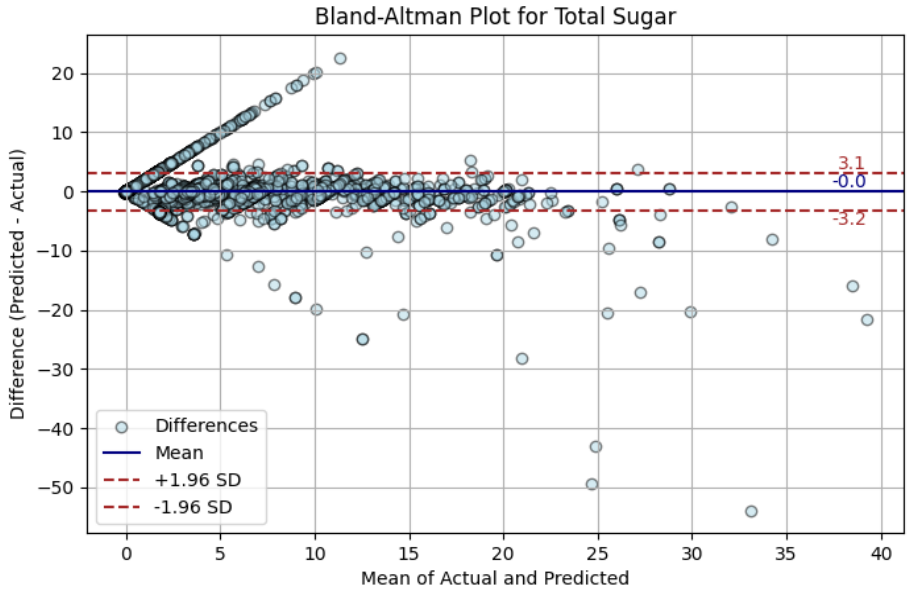

Figure 13: Agreement of Total Sugar Prediction vs Actual

Figure 13 follows the same theme as Figure 12. We can also predict that the majority of the predictions fall within the agreement limits. Again, prediction errors appear at higher sugar levels, and the mean difference between predicted and actual values close to zero, indicating little to no systematic error. It displays more outliers but with, once more, no clear tend to increasing or decreasing error.

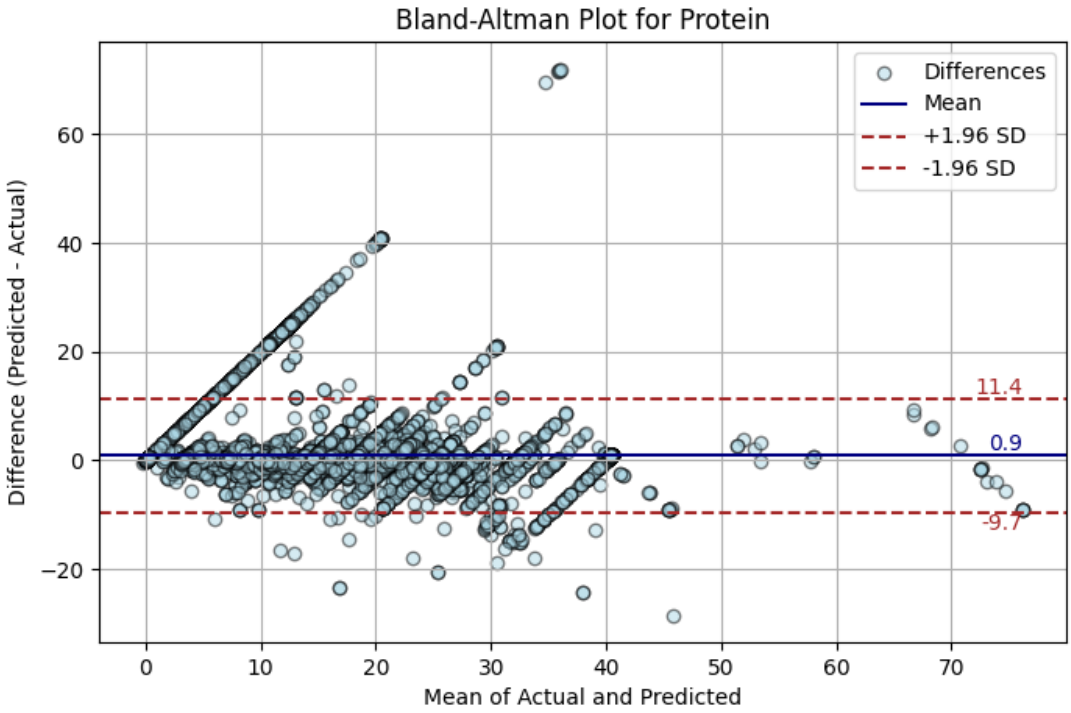

Figure 14: Agreement of Protein Prediction vs Actual

The Bland-Altman plot, in Figure 14, demonstrates similar results as Figure 12 and 13. Majority of the prediction are grouped in the agreement limits, we don't see the tendency to have prediction errors in higher values, and it manages to have even less outliers than the other two.

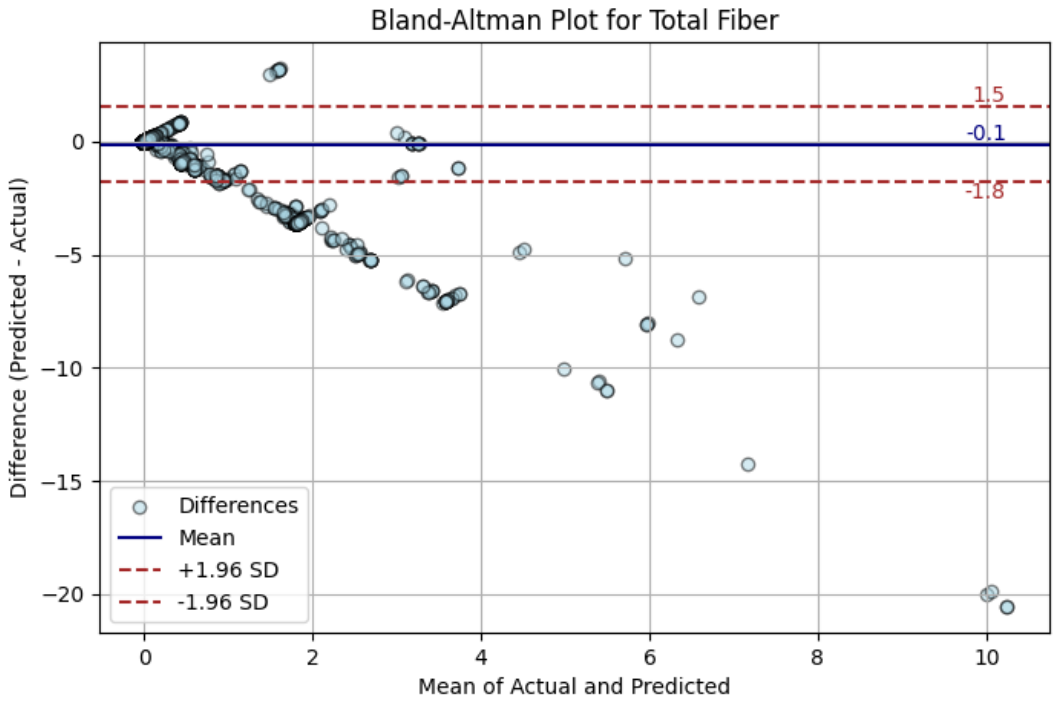

Figure 15: Agreement of Total Fiber Prediction vs Actual

On the contrary, and closer to the results of Figure 11, in Figure 15 we see a wider spread of prediction errors since most data points are outside the limits of agreement. Visible systematic bias is displayed when the model predicts this nutrient, highlighting its difficulty in doing so.

4.4.3 Model Evaluation Using Scatter Plots

In this section, we will also discuss the strengths and weaknesses of the model when it tries to predict each specific nutrient, by visualizing the results on scatter plots. The results are produced from the output of the validation set, of the data sample, with configuration of Table 12. Analysis of the Scatter plots helped us to gather additional information on the generalizability and accuracy of the model, for each specific nutrient. A Scatter plot, visualizes the actual agreement between the predicted and actual nutrient value, providing a detailed picture of both the model's accuracy and its variance in prediction.

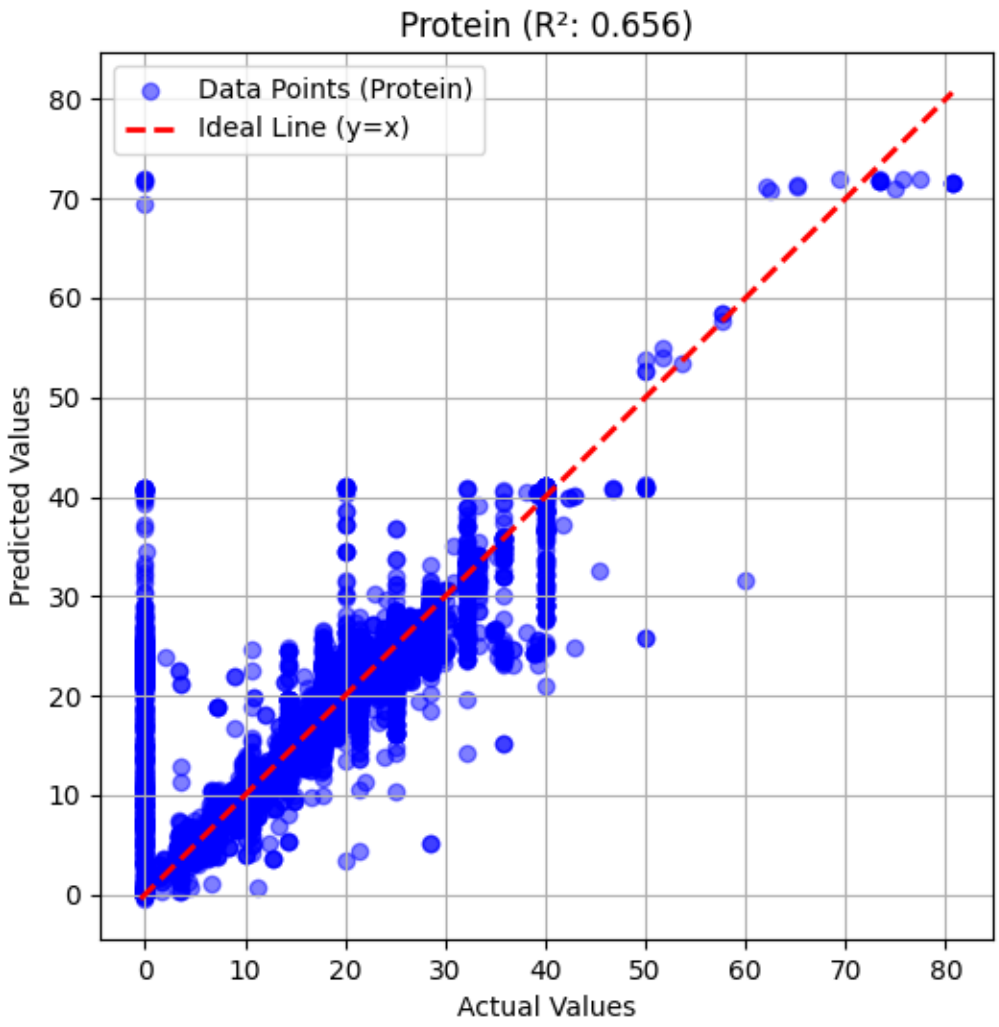

Figure 16: Comparison plot of the Predicted vs Actual values for Protein

The model shows a moderate correlation between actual and predicted protein values since most of the values are pretty close to diagonal, with some dispersion around it. It makes it clear that the model is weak in predicting 0 values, but there are no other obvious patterns of over- or under-estimation.

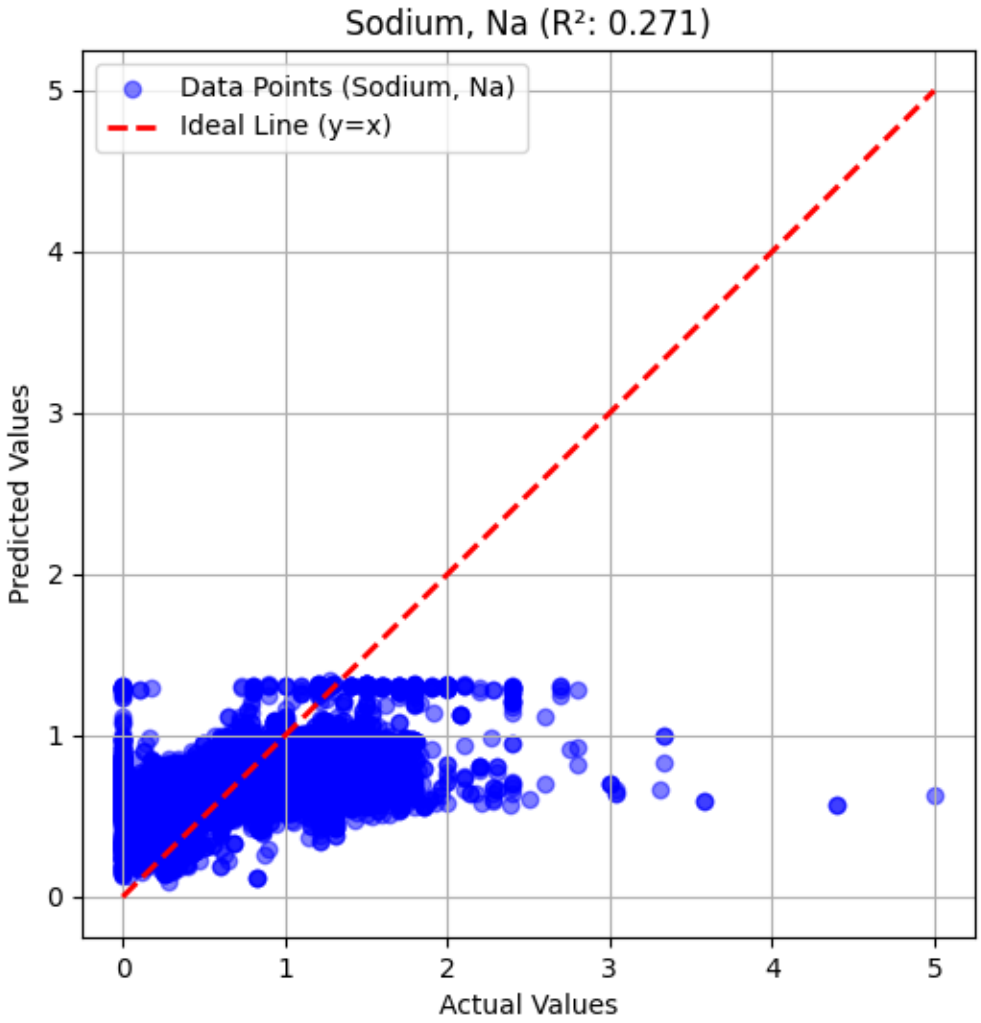

Figure 17: Comparison plot of the Predicted vs Actual values for Sodium

The model struggles to predict sodium values accurately, with a low R^2 score indicating weak correlation. Most of the predictions cluster in low values, showcasing a systematic bias where the model seems to underestimate the values.

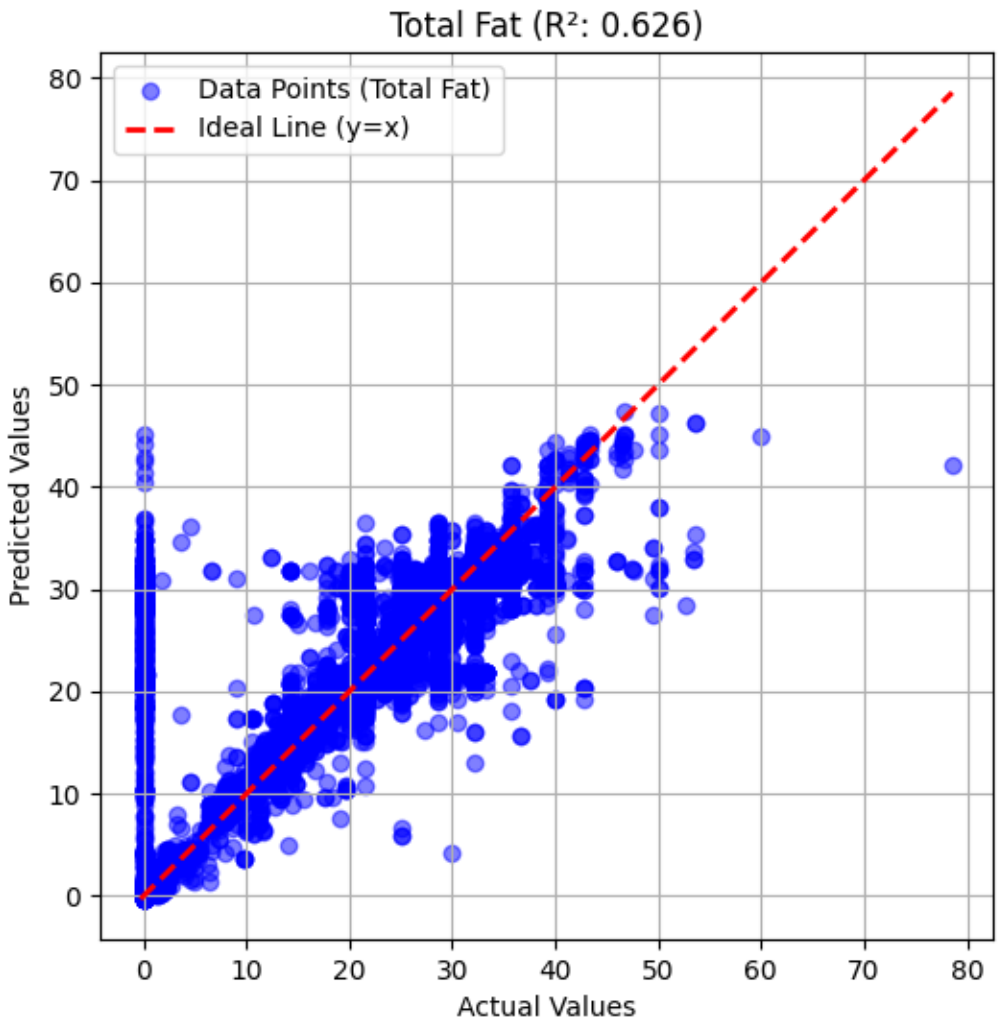

Figure 18: Comparison plot of the Predicted vs Actual values for Total Fat

The model demonstrates a moderate predictive performance for total fat, with an R2 score of 0.626, much similar to its protein prediction performance. While many predictions align with the ideal line, we can observe that as the values get higher the model seems to under-estimate, but not with a clear indication of systematic bias. It is highly visible again that the model does not manage to recognize 0 values.

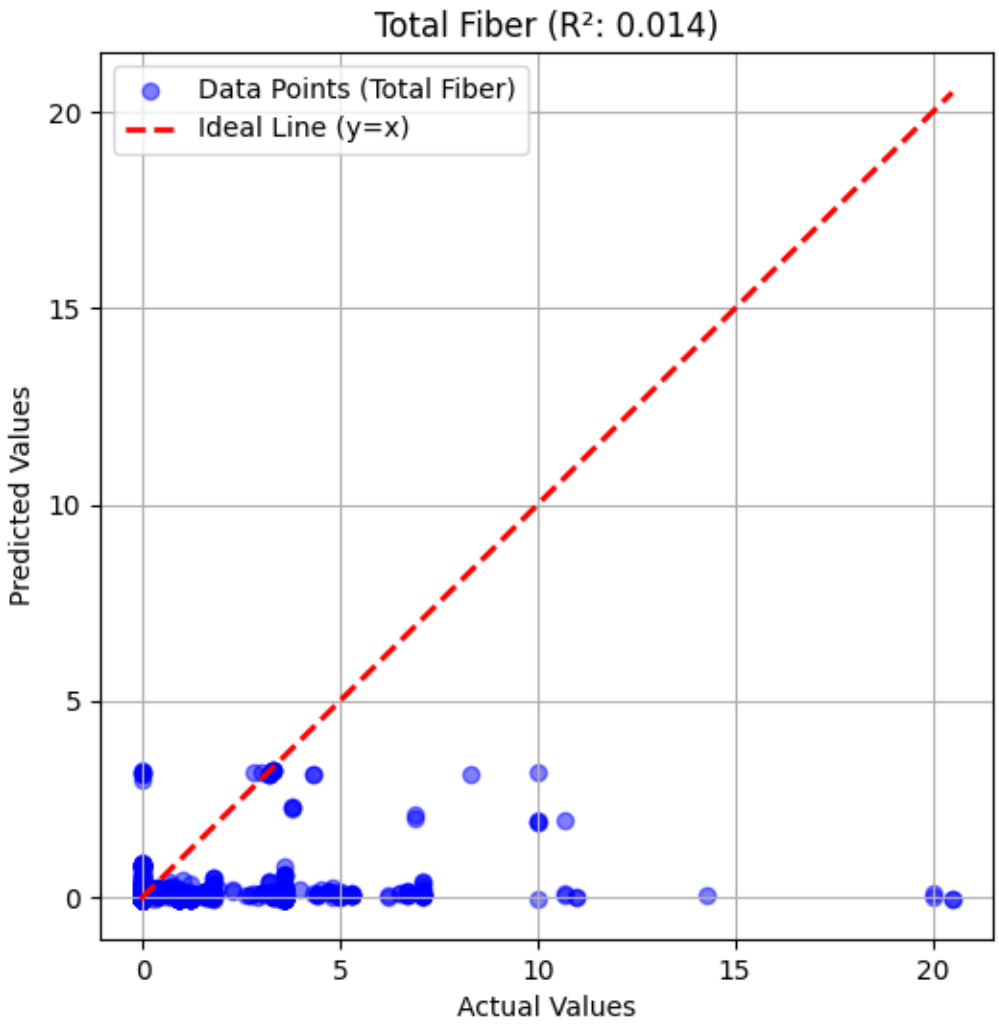

Figure 19: Comparison plot of the Predicted vs Actual values for Total Fiber

The model exhibits weak predictive performance for total fiber — if not non-existent. Predictions cluster near zero, indicating a failure to capture variance, possibly due to imbalanced training data or feature limitations. The model underestimates the values as it did when predicting Sodium, making it clear that it doesn't have predictive abilities for those two.

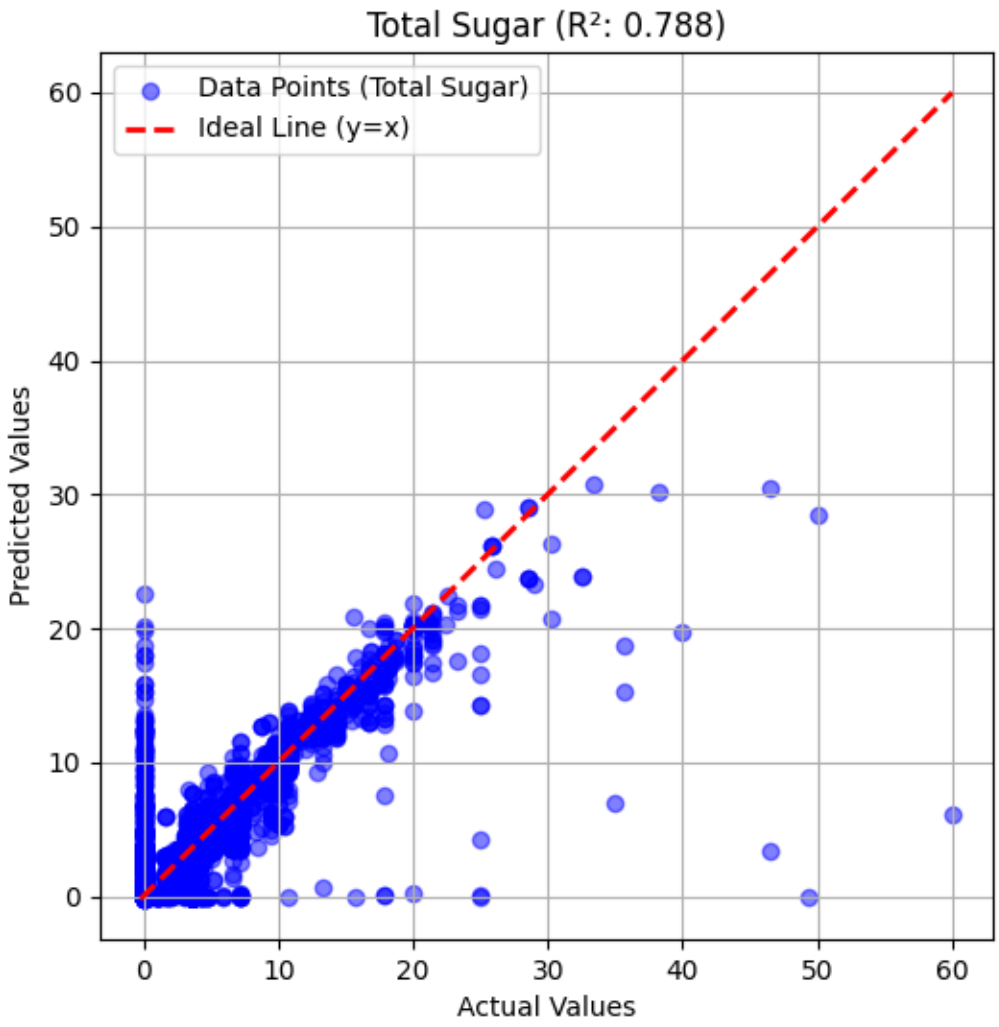

Figure 20: Comparison plot of the Predicted vs Actual values for Total Sugar

The model demonstrates a solid predictive capability for total sugar, with an R2 score of 0.788. Predictions align well with actual values, making Total Sugar the strongest feature of our model. It is again clear that the model struggles to recognize 0 values, and it systematically underestimates as the values get higher.

4.4.4 Best Performing Models and Final Selection

The comparative analysis of training and validation loss, combined with different model configurations, proved to be fundamental for optimizing predictive performance. Such comparisons guided us to base our models on SmoothL1Loss, while choosing the loss function, since it consistently outperformed those using MSELoss, L1Loss, or HuberLoss, achieving lower final validation loss and higher R^2 scores (Pillay et al., 2023; Hung et al., 2024). An interesting point of the study is that the models perform better when trained on structured and targeted data, which comes in contrast with the belief that very sets increase the model's performance since they provide more data. It seems that the quality and organization of the data plays a greater role than the quantity (Ma et al., 2022).

On discussions about predictive accuracy versus computational cost, the results reveal that more complex and deeper networks achieved marginally better validation values, but at much higher training time and hardware resource requirements (Menghani, 2023; Keremidchiev, 2024). It is safe to say that after a point, adding more layers and parameters does not always provide a significant benefit on the model's performance. In contrast, normalization techniques such as dropout and batch normalization achieved similar accuracy at much lower computational cost, lighting the way to develop structured data with simplified architectures to achieve reliable predictions without excessive resource consumption.

In order to choose the right model for any specific application, we have to take into account accuracy, performance, and generalization ability. As displayed in this study, when finding a balance between loss function, appropriate data preprocessing, and medium-complexity architectures — models are capable of being robust without exhausting computational resources (Menghani, 2023). Without saying that deeper neural networks can offer better performance, the cost of time and hardware requirements make them less practical for large-scale or real-time applications (Marculescu et al., 2018; Bartoldson et al., 2023).

The next challenge is to optimize accuracy while maintaining efficiency, through more flexible loss functions and targeted feature selection — especially for real-time applications — making AI effective in the food sector balancing between computational economy and practical reliability (Bui et al., 2017; Vijayakumar & Bharathi, 2024).

Chapter 5
Discussion & Conclusions

The findings in the study align with the hypothesis that category-specific datasets enhance predictive stability, that was previously confirmed in the research on domain-adaptive learning (Zhang et al., 2020). Based on the results, we conclude that focused datasets — category filtered data, cheese with 80,000 entries — achieve lower validation loss compared to larger — non-filtered general datasets of 300,000 entries. Based on other studies, we clearly see that the result aligns, emphasizing this way the advantages of specialized embeddings in understanding domain-relevant patterns (Budach et al., 2022), with the increase in validation loss for generalized datasets indicating that an excess of heterogeneous data introduce noise, negatively affecting model performance and potential (Arjovsky, 2020; Barbiero et al., 2020). The above highlight the caution in which the model generalization should be approached, ensuring that extracted features maintain high discriminative power (Ferrao et al., 2016; Alexandropoulos et al., 2019; Hassler et al., 2019). Consequently, dataset structuring is critical for improving classification robustness and mitigating learning biases,as previously observed in predictive modeling studies (Ferrao et al., 2016; Van Giffen et al., 2022).

The metrics, used in the evaluation of model performance, revealed how the choice of loss function and training configurations affect the predictive accuracy and stability. Through the experimentation, we managed to highlight the performance superiority of SmoothL1Loss when comparing it to other common loss functions, since it effectively managed to balance error penalization and maintained model robustness (Ciampiconi et al., 2023). Unlike the common loss functions — MSELoss and L1Loss — which mostly tend to either excessively penalize large errors or underestimate minor deviations, SmoothL1Loss demonstrates an adaptive nature, since it combines the best of both worlds, minimizing overfitting while improving generalization. An adaptability that is particularly valuable in large datasets, where extreme values can disproportionately influence predictions (Jadon et al., 2024).

In the study, is clearly stated, the importance of structured and refined data. The stage of preparation impacts the model's performance and ability to generalize, rather than the dataset volume. As mentioned, the smaller, structured dataset, had the ability to generalize more

accurately, in contrary to the unorganized one, where the model's ability to generalize dropped as the dataset was getting bigger, with the largest showing hints of memorization (Ma et al., 2022). Data that refer to a specific category or follow the same patterns, help the model to avoid distractions as they don't introduce noise. Without wanting to underestimate the sample size importance, we want to highlight the results a clear and structured dataset can provide. We need to always opt for both, large and structured dataset, but if we need to pick one, the cleaner environment that the structured one provides will be the winning choice (Zhu et al., 2023). Complex models can capture more complex relationships in the data — that doesn't necessarily mean they generalize better — that highlights the importance to find balance between predictive accuracy and computational efficiency (Menghani, 2023). Excessive complexity doesn't guarantee better results, on the contrary, it can lead to lag, overfitting, and loss of transparency.

As analyzed in the research, other methods than the traditional laboratory analysis, can provide an accurate overview of the food nutritional composition, like NLP based technologies, that can predict a food's nutritional composition right out of the box, by "reading" the ingredient list of the branded food product. The current methods to achieve those results are mainly based on laboratory analysis, which provide extremely accurate results, but comes with the cost of both time and money, making them unviable to apply as a widespread application (Hawley et al., 2013; Temple, 2020). The proposed approach speaks of a computational mechanism that bridges the gap between strings of ingredients composing the food and its numerical nutritional data, enabling the automatic extraction of information related to its ingredients. By using advanced machine learning algorithms, and in particular transformers such as DistilBERT, the present study aims to set a new precedent for analyzing food components and associating them with accurate nutritional values (Kasapila & Shaarani, 2011). This contribution is not limited to automation alone but seeks to improve the transparency, accessibility, and standardization of nutritional information, addressing gaps in both regulatory compliance and consumer understanding.

The research's innovation and significance implications are its methods integrating natural language processing with nutrition predictive modeling. Unlike current approaches with fixed databases and preset hierarchical structures, the considered algorithms formulate self-adaptive algorithms that can enhance with more data exposure. The study offers an interpretable

framework that is able to generate predictive estimates while ensuring adherence to various compliance boundaries. Given the intricacy of the global food industry where regulatory demands and market tastes are in perpetual flux, this kind of flexibility is useful (Nayak & Waterson, 2019; Cohen & Kouvelis, 2021). This research resolves inefficiencies in food labeling classification and multi-jurisdictional validation by designing a system that automates unified food labeling practices transversal to jurisdictions.

A deeper interpretation of the contribution of this research is that it makes access to nutritional information more democratic and accessible to different socioeconomic strata. Existing food labelling systems often favor regions with strong regulatory frameworks, leaving less developed countries with limited oversight and control. The ability to computationally extract nutritional values from ingredient lists enables the protection of public health, especially in areas where laboratory analysis is impractical, providing the ability to the end-user or consumer of the product, to have a better overview of the product. This reduces the reliance on self-reporting by manufacturers, minimizing the risk of false nutritional claims, and provides the ability to the manufacturer to test recipes for new products, without having to run laboratory analysis for each new recipe idea, minimizing the cost of production. It is also worth mentioning the promotion of transparency and ethics in product labeling, providing to both consumers and regulators a fast and accurate way to verify nutritional claims.

The overall impact of the given research is that manages to improve the accuracy of food labeling, by contributing to the creation of a more accessible and fair system for providing nutritional information, without leaving the traditional and accurate methods aside, but working together, with the artificial intelligence as a guide that provides a prediction of new product or raises caution on a product right on the self of the store, and the laboratory analysis playing the role of the reviewer that accurately tests and approves the cases accordingly, we demonstrate the enormous potential of artificial intelligence in the field of nutrition and public health.

Conclusions

When testing various loss functions in this study, SmoothL1Loss clearly stood out, since it successfully managed to balance stability and accuracy, making it more than efficient for fine-tuning the model. In contrary to classic loss functions, that overemphasize large errors or ignore small — but important — deviations. SmoothL1Loss offers both handling of outliers and maintained sensitivity to detail, allowing the model to adapt to different types of ingredient lists and make accurate predictions, regardless of the dataset size or categorization. It is also computationally efficient, meaning you don't have to sacrifice speed for accuracy. Overall, its use lays the foundation for more reliable predictive models.

However, just as important as the loss function, is the way the dataset is organized and cleaned. Investing time in filtering and curating your training data yields tangible results. A well-organized dataset allows the model to more easily identify the right patterns, reduces the noise from low quality or unclear ingredient lists, and leads to more reliable results. Almost always, a well-structured smaller dataset outperforms a large and chaotic one: it preserves the meaningful relationships between features and avoids the learning being corrupted by random or irrelevant data. This not only improves the accuracy of predictions but also makes the model easier to interpret and more adaptable to real-world applications, such as food labeling.

Given the significance of data preprocessing, we conclude that adding layers or parameters doesn't necessarily guarantee better results — it often hurts performance and increases the risk of overfitting. Keeping a simple, focused architecture, focusing on the important features and using as much deep learning as necessary, makes the system lighter, faster, and more reliable. When there is balance between feature engineering and efficiency, the model becomes easier to implement and maintain — whether for regulatory use or everyday consumer support.

Success seems to depend on three key factors when suggesting deep learning for nutritional data analysis:

- Consistency and plenty of data,
- Avoid unwanted noise of empty or misguiding data,
- Focus on real-world applicability.

The next step in model development should emphasize better data preprocessing, flexible loss functions such as SmoothL1Loss, and avoiding unnecessary complexity. Fine-tuning all these elements — with an emphasis on future scalability to meet new regulatory requirements and changing consumer needs — is key to developing effective, reliable, and sustainable predictive systems in nutritional science.

Future Work

Looking ahead, it is clear that prediction with deep learning models requires more than just standard loss functions. The next step is to develop more flexible and adaptive loss mechanisms (hybrid loss mechanisms), which can dynamically adapt to the complexity and idiosyncrasies of each data set. By combining the advantages of different loss functions, a model can be created that “knows” when to impose severe penalties for large errors, but also when to reward small but significant deviations. Thus, instead of applying a single approach, hybrid schemes can adapt their behavior depending on what really matters for the quality of predictions. This will allow models to intelligently deal with outliers while remaining sensitive to critical details. Such an approach is particularly useful when ingredient lists are heterogeneous or the distribution of nutritional values varies significantly. From this research comes the need to combine data handling optimization strategies with AI-tools, increasing the likelihood of this becoming a reality.

However, as we identified, the quality, quantity and structure of the dataset play an important role to a model’s capability in generalizing. With manual curation being a painful step in data preparation, since is time-consuming, subjective, and unsustainable on large-scale datasets. Automated feature selection emerges to open a world where the system could preserve key features and discard unnecessary ones, and continuously adapt to new ingredient compositions.

To be applicable in real life, models need to be more than just accurate — they need to be fast and understandable. With our outmost attention to the efficiency of neural networks, through smart changes in architecture, parallel processing, and optimization of parameter management, to achieve fast predictions without losing accuracy. The food sector and Artificial Intelligence technologies will only be able to work together when neural networks become capable to adapt

to different datasets, make efficient use of information, and offer simplicity without sacrificing quality.

With fast and flexible models, able to keep up with the changing demands of food science and everyday life, the doors open to a future where predictive solutions will be integrated into the daily lives of regulators, industry, and consumers.

References

Adolfo, C. M. S., Chizari, H., Win, T. Y., & Al-Majeed, S. (2021). Sample reduction for physiological data analysis using principal component analysis in artificial neural network. *Applied Sciences, 11*(17), 8240. (http://dx.doi.org/10.3390/app11178240)

Afshar, P., Mohammadi, A., Plataniotis, K. N., Oikonomou, A., & Benali, H. (2019). From handcrafted to deep-learning-based cancer radiomics: Challenges and opportunities. *IEEE Signal Processing Magazine, 36*(4), 132-160. (https://doi.org/10.1109/MSP.2019.2900993)

Ahmed, T., Wijewardane, N. K., Lu, Y., Jones, D. S., Kudenov, M., Williams, C., ... & Kamruzzaman, M. (2024). Advancing sweetpotato quality assessment with hyperspectral imaging and explainable artificial intelligence. *Computers and Electronics in Agriculture, 220*, 108855. (10.1016/j.compag.2024.108855)

Akbari, M. (2023). Recipe popularity prediction in Finnish social media by machine learning models (Master's thesis, M. Akbari). (https://oulurepo.oulu.fi/bitstream/handle/10024/43040/nbnfioulu-202310133120.pdf?sequence=1)

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S. and Vollgraf, R., 2019, June. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)* (pp. 54-59). (https://doi.org/10.18653/v1/N19-4010)

Alexandropoulos, S. A. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. *The Knowledge Engineering Review, 34*, e1. (https://doi.org/10.1017/S026988891800036X)

Al-Huthaifi, R., Li, T., Al-Huda, Z., Huang, W., Luo, Z., & Xie, P. (2024). FedGODE: Secure traffic flow prediction based on federated learning and graph ordinary differential equation networks. *Knowledge-Based Systems, 299*, 112029. (https://doi.org/10.1016/j.knosys.2024.112029)

Areta, Ourania & Yalcinkaya, Elmas. (2024). A Comparative Analysis of Machine Learning Models for Time Prediction in Food Delivery Operations. 4. 43-56. (ISBN: 978-605-69730-2-4)

Arjovsky, M. (2020). Out of distribution generalization in machine learning (Doctoral dissertation, New York University). (https://doi.org/10.48550/arXiv.2103.02667)

Armand, T. P. T., Nfor, K. A., Kim, J. I., & Kim, H. C. (2024). Applications of artificial intelligence, machine learning, and deep learning in nutrition: A systematic review. *Nutrients, 16*(7), 1073. (https://doi.org/10.3390/nu16071073)

Ayman, A., Mansour, Y., & Eldaly, H. (2024). Applying machine learning algorithms to architectural parameters for form generation. *Automation in Construction, 166*, 105624. (https://doi.org/10.1016/j.autcon.2024.105624)

Bacarella, S., Altamore, L., Valdesi, V., Chironi, S. and Ingrassia, M., 2015. Importance of food labelling as a means of information and traceability according to consumers. *Advances in Horticultural Science*, *29*(2/3), pp.145-151. (http://dx.doi.org/10.13128/ahs-22695)

Baer, J. (2015). The importance of domain-specific expertise in creativity. *Roeper Review, 37*(3), 165–178. (https://doi.org/10.1080/02783193.2015.1047480)

Barbiero, P., Squillero, G., & Tonda, A. (2020). Modeling generalization in machine learning: A methodological and computational study. (https://doi.org/10.48550/arXiv.2006.15680)

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. (https://doi.org/10.48550/arXiv.2004.05150)

Berryman, P. ed., 2014. *Advances in Food and Beverage Labelling: Information and Regulations*. Elsevier. (ISBN 13: 9781782420859)

Bartoldson, B. R., Kailkhura, B., & Blalock, D. (2023). Compute-efficient deep learning: Algorithmic trends and opportunities. *Journal of Machine Learning Research, 24*(122), 1-77. (https://doi.org/10.48550/arXiv.2210.06640)

Bhowmik, R., 2021. *Neural Methods for Entity-Centric Knowledge Extraction and Reasoning in Natural Language* (Doctoral dissertation, Rutgers The State University of New Jersey, School of Graduate Studies). (https://doi.org/doi:10.7282/t3-gk5z-3g02)

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). Feature selection for high-dimensional data. *Progress in Artificial Intelligence, 5*, 65-75. (http://dx.doi.org/10.1007/978-3-319-21858-8)

Bottou, L., Curtis, F. E., & Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review, 60*(2), 223-311. (https://doi.org/10.48550/arXiv.1606.04838)

Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., ... & Harmouch, H. (2022). The effects of data quality on machine learning performance. *arXiv preprint arXiv:2207.14529*. (https://doi.org/10.1016/j.is.2025.102549)

Bui, N., Cesana, M., Hosseini, S. A., Liao, Q., Malanchini, I., & Widmer, J. (2017). A survey of anticipatory mobile networking: Context-based classification, prediction methodologies, and optimization techniques. *IEEE Communications Surveys & Tutorials, 19*(3), 1790-1821. (http://dx.doi.org/10.1109/COMST.2017.2694140)

Buolamwini, J. and Gebru, T., 2018, January. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR 81:77-91 (https://proceedings.mlr.press/v81/buolamwini18a.html)

Cao, R., Li, J., Ding, H., Zhao, T., Guo, Z., Li, Y., ... & Qiu, J. (2024). Synergistic approaches of AI and NMR in enhancing food component analysis: A comprehensive review. *Trends in Food Science & Technology*, 104852. (https://doi.org/10.1016/j.tifs.2024.104852)

Chan, L. E. (2023). Semantic evaluation of environmental & nutrition factors impacting female reproductive disorders. (https://doi.org/10.1016/j.ijmedinf.2024.105461)

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16-28. (https://doi.org/10.1016/j.compeleceng.2013.11.024)

Chatterjee, S., & Zielinski, P. (2022). On the generalization mystery in deep learning. (https://doi.org/10.48550/arXiv.2203.10036)

Chattopadhyay, P., Balaji, Y., & Hoffman, J. (2020). Learning to balance specificity and invariance for in and out of domain generalization. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part IX 16* (pp. 301–318). Springer International Publishing. (https://doi.org/10.48550/arXiv.2008.12839)

Cheng, S., Chen, W., Liu, W. et al. Dynamic training for handling textual label noise. Appl Intell 54, 11161–11176 (2024). (https://doi.org/10.1007/s10489-024-05738-x)

Chew, N.W., Ng, C.H., Tan, D.J.H., Kong, G., Lin, C., Chin, Y.H., Lim, W.H., Huang, D.Q., Quek, J., Fu, C.E. and Xiao, J., 2023. The global burden of metabolic disease: Data from 2000 to 2019. *Cell Metabolism*, *35*(3), pp.414-428. (https://doi.org/10.1016/j.cmet.2023.02.003)

Choudhury, A. (2025). Data collection and preprocessing for generative AI. In *Generative AI for Business Analytics and Strategic Decision Making in Service Industry* (pp. 1–32). IGI Global Scientific Publishing. (http://dx.doi.org/10.4018/979-8-3693-7026-1.ch001)

Chowdhary, K.R. (2020). Natural Language Processing. In: Fundamentals of Artificial Intelligence. Springer, New Delhi. (https://doi.org/10.1007/978-81-322-3972-7_19)

Ciampiconi, L., Elwood, A., Leonardi, M., Mohamed, A., & Rozza, A. (2023). A survey and taxonomy of loss functions in machine learning. (https://doi.org/10.48550/arXiv.2301.05579)

Cohen, M. A., & Kouvelis, P. (2021). Revisit of AAA excellence of global value chains: Robustness, resilience, and realignment. *Production and Operations Management, 30(3),* 633-643. (http://dx.doi.org/10.1111/poms.13305)

Conneau, A., 2019. Unsupervised cross-lingual representation learning at scale. (https://doi.org/10.48550/arXiv.1911.02116)

Cui, Y., Song, Y., Sun, C., Howard, A., & Belongie, S. (2018). Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition (pp. 4109–4118). (https://doi.org/10.48550/arXiv.1806.06193)

Davis, J., Bransen, L., Devos, L., Jaspers, A., Meert, W., Robberechts, P., ... & Van Roy, M. (2024). Methodology and evaluation in sports analytics: challenges, approaches, and lessons learned. *Machine Learning, 113*(9), 6977-7010. (http://dx.doi.org/10.1007/s10994-024-06585-0)

Delanerolle, G., Yang, X., Shetty, S., Raymont, V., Shetty, A., Phiri, P., ... & Shi, J. Q. (2021). Artificial intelligence: A rapid case for advancement in the personalization of gynaecology/obstetric and mental health care. *Women's Health, 17*, 17455065211018111. (https://doi.org/10.1177/17455065211018111)

Delfani, P., Thuraga, V., Banerjee, B., & Chawade, A. (2024). Integrative approaches in modern agriculture: IoT, ML and AI for disease forecasting amidst climate change. *Precision Agriculture, 25*(5), 2589-2613. (http://dx.doi.org/10.1007/s11119-024-10164-7)

Delgado-Lista, J., Alcala-Diaz, J.F., Torres-Peña, J.D., Quintana-Navarro, G.M., Fuentes, F., Garcia-Rios, A., Ortiz-Morales, A.M., Gonzalez-Requero, A.I., Perez-Caballero, A.I., Yubero-Serrano, E.M. and Rangel-Zuñiga, O.A., 2022. Long-term secondary prevention of cardiovascular disease with a Mediterranean diet and a low-fat diet (CORDIOPREV): a randomised controlled trial. *The Lancet*, *399*(10338), pp.1876-1885. (https://doi.org/10.1016/s0140-6736(22)00122-2)

Delobelle, P., Winters, T. and Berendt, B., 2020. Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*. (https://doi.org/10.18653/v1/2020.findings-emnlp.292)

Dessain, J. (2022). Machine learning models predicting returns: Why most popular performance metrics are misleading and proposal for an efficient metric. *Expert Systems with Applications, 199*, 116970. (https://doi.org/10.1016/j.eswa.2022.116970)

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language*

technologies, volume 1 (long and short papers) (pp. 4171–4186). (https://doi.org/10.18653/v1/N19-1423)

Díaz, L. D., Fernández-Ruiz, V., & Cámara, M. (2020). An international regulatory review of food health-related claims in functional food products labelling. *Journal of Functional Foods*, *68*, 103896 (https://doi.org/10.1016/j.jff.2020.103896)

Dinkel, H., Wu, M., & Yu, K. (2021). Towards duration robust weakly supervised sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 887-900. (https://doi.org/10.1109/TASLP.2021.3054313)

Do, K., Nguyen, M. D., Hoa, N. T., Tran-Thanh, L., Tran, N. H., & Pham, Q. V. (2024). Revisiting LARS for large batch training generalization of neural networks. *IEEE Transactions on Artificial Intelligence*. (https://doi.org/10.48550/arXiv.2309.14053)

Ennaji, O., Vergütz, L., & El Allali, A. (2023). Machine learning in nutrient management: A review. *Artificial Intelligence in Agriculture, 9*, 1-11. (http://dx.doi.org/10.1016/j.aiia.2023.06.001)

European Commission., 2019. Ethics Guidelines for Trustworthy AI. Available at: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf .

Faber, K., Corizzo, R., Sniezynski, B., & Japkowicz, N. (2024). Lifelong continual learning for anomaly detection: New challenges, perspectives, and insights. *IEEE Access, 12*, 41364-41380. (https://doi.org/10.1109/ACCESS.2024.3377690)

Facioni, M.S., Raspini, B., Pivari, F. et al. Nutritional management of lactose intolerance: the importance of diet and food labelling. J Transl Med 18, 260 (2020). (https://doi.org/10.1186/s12967-020-02429-2)

Ferrao, J. C., Oliveira, M. D., Janela, F., & Martins, H. M. (2016). Preprocessing structured clinical data for predictive modeling and decision support. *Applied Clinical Informatics, 7*(4), 1135–1153. (https://doi.org/10.4338/aci-2016-03-soa-0035)

Freiesleben, T., Grote, T. Beyond generalization: a theory of robustness in machine learning. Synthese 202, 109 (2023). (https://doi.org/10.1007/s11229-023-04334-9)

Folorunso, O., Ojo, O., Busari, M., Adebayo, M., Joshua, A., Folorunso, D., ... & Olabanjo, O. (2023). Exploring machine learning models for soil nutrient properties prediction: A systematic review. *Big Data and Cognitive Computing, 7*(2), 113. (https://doi.org/10.3390/bdcc7020113

Ghai, S., Shrivastava, R., & Jain, S. (2024). Analysis of *Mycobacterium Fortuitum* proteome using machine learning techniques (Doctoral dissertation, Jaypee University of Information Technology, Solan, HP). (http://ir.juit.ac.in:8080/jspui/jspui/handle/123456789/11246)

Gholizadeh-Moghaddam, M., Shahdadian, F., Shirani, F., Hadi, A., Clark, C.C. and Rouhani, M.H., 2023. The effect of a low versus high sodium diet on blood pressure in diabetic patients: A systematic review and meta-analysis of clinical trials. *Food Science & Nutrition*, *11*(4), pp.1622-1633. (http://dx.doi.org/10.1002/fsn3.3212)

García, J., Leiva-Araos, A., Diaz-Saavedra, E., Moraga, P., Pinto, H., & Yepes, V. (2023). Relevance of machine learning techniques in water infrastructure integrity and quality: A review powered by natural language processing. *Applied Sciences, 13*(22), 12497. (https://doi.org/10.3390/app132212497)

Garg, N., Dwivedi, P. A Novel Approach for Exploring Data-Driven Nutritional Insights Using Clustering and Dimensionality Reduction Techniques. SN COMPUT. SCI. 5, 1019 (2024). (https://doi.org/10.1007/s42979-024-03397-w)

Gawusu, S., Jamatutu, S. A., Zhang, X., Moomin, S. T., Ahmed, A., Mensah, R. A., ... & Ackah, I. (2024). Spatial analysis and predictive modeling of energy poverty: Insights for policy implementation. *Environment, Development and Sustainability*, 1-48. (https://doi.org/10.1007/s10668-024-05015-4)

Geiping, J., Goldblum, M., Pope, P. E., Moeller, M., & Goldstein, T. (2021). Stochastic training is not necessary for generalization. (https://doi.org/10.48550/arXiv.2109.14119)

Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology, 162*, 107268. (https://doi.org/10.1016/j.infsof.2023.107268)

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. (ISBN: 9780262035613)

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., ... & He, K. (2017). Accurate, large minibatch SGD: Training ImageNet in 1 hour. (https://doi.org/10.48550/arXiv.1706.02677)

Harris, E., 2024. Ultraprocessed Foods Linked With 32 Types of Health Problems. *JAMA*. 331(15):1265. doi:10.1001/jama.2024.2088 (https://doi.org/10.1001/jama.2024.2088)

Hassler, A. P., Menasalvas, E., García-García, F. J., Rodríguez-Mañas, L., & Holzinger, A. (2019). Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome. BMC Medical Informatics and Decision Making, 19, 1–17. (https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0747-6)

Hawley, K. L., Roberto, C. A., Bragg, M. A., Liu, P. J., Schwartz, M. B., & Brownell, K. D. (2013). The science on front-of-package food labels. *Public health nutrition, 16(3),* 430-439. (http://dx.doi.org/10.1017/S1368980012000754)

He, B., Noci, L., Paliotta, D., Schlag, I., & Hofmann, T. (2025). Understanding and minimizing outlier features in transformer training. *Advances in Neural Information Processing Systems, 37*, 83786-83846. (https://doi.org/10.48550/arXiv.2405.19279)

Hemdev, P., 2009. *Information Extraction: A Smart Calendar Application* (Doctoral dissertation, University of Oxford). (ISBN-13: 978-3639353051)

Henderikx, F., 2017. Labelling of food: a challenge for many. *Veterinarski glasnik*, *71*(1), pp.16-23. (https://doi.org/10.2298/VETGL170214001H)

Hill, A.J., Basourakos, S.P., Lewicki, P., Wu, X., Arenas-Gallo, C., Chuang, D., Bodner, D., Jaeger, I., Nevo, A., Zell, M. and Markt, S.C., 2022. Incidence of kidney stones in the United States: the

continuous national health and nutrition examination survey. *The Journal of urology*, *207*(4), pp.851-856. (http://dx.doi.org/10.1097/JU.0000000000002331)

Hoffer, E., Hubara, I., & Soudry, D. (2017). Train longer, generalize better: Closing the generalization gap in large batch training. *Advances in Neural Information Processing Systems, 30*. (https://doi.org/10.48550/arXiv.1705.08741)

Holland, Sarah & Newman, Sarah & Joseph, Joshua & Chmielinski, Kasia. (2020). The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards. (http://dx.doi.org/10.5040/9781509932771.ch-001)

Hosamo, H., & Mazzetto, S. (2024). Performance evaluation of machine learning models for predicting energy consumption and occupant dissatisfaction in buildings. *Buildings, 15*(1), 39. (https://doi.org/10.3390/buildings15010039)

Hosseinpour-Niazi, S., Niknam, M., Amiri, P., Mirmiran, P., Ainy, E., Izadi, N., Gaeini, Z. and Azizi, F., 2024. The association between ultra-processed food consumption and health-related quality of life differs across lifestyle and socioeconomic strata. *BMC public health*, *24*(1), p.1955. (https://doi.org/10.1186/s12889-024-20189-2)

Hu, G., Ahmed, M. and L'Abbé, M.R., 2023. Natural language processing and machine learning approaches for food categorization and nutrition quality prediction compared with traditional methods. *The American Journal of Clinical Nutrition*, *117*(3), pp.553-563. (https://doi.org/10.1016/j.ajcnut.2022.11.022)

Hua, A., Dhaliwal, M.P., Burke, R., Pullela, L. and Qin, Y., 2024. NutriBench: A Dataset for Evaluating Large Language Models on Nutrition Estimation from Meal Descriptions. *arXiv preprint arXiv:2407.12843*. (https://doi.org/10.48550/arXiv.2407.12843)

Hung, J. W., Huang, P. C., & Li, L. Y. (2024). Employing Huber and TAP Losses to Improve Inter-SubNet in Speech Enhancement. *Future Internet, 16*(10), 360. (http://dx.doi.org/10.3390/fi16100360)

Ibrahim, A., Khodadadi, E., Khodadadi, E., Dutta, P. K., Bailek, N., & Abdelhamid, A. A. (2024). Apple perfection: Assessing apple quality with waterwheel plant algorithm for feature selection and logistic regression for classification. *Journal of Artificial Intelligence in Engineering Practice, 1*(1), 34-48. (https://dx.doi.org/10.21608/jaiep.2024.355003)

Ikram, A., & Aslam, W. (2024). Enhancing intercropping yield predictability using optimally driven feedback neural network and loss functions. *IEEE Access*. (http://dx.doi.org/10.1109/ACCESS.2024.3486101)

Ishwarya, V. S., & Kothandaraman, M. (2024). A novel feature-fusion-based sparse masked attention network for acoustic echo cancellation using wavelet and STFT synergies. *Circuits, Systems, and Signal Processing*, 1-20. (http://dx.doi.org/10.1007/s00034-024-02955-0)

Ispirova, G., Eftimov, T., & Koroušić Seljak, B. (2020). P-nut: Predicting nutrient content from short text descriptions. *Mathematics, 8*(10), 1811. (https://doi.org/10.3390/math8101811)

Ispirova, Gordana. (2022). Exploiting Domain Knowledge in Predictive Learning from Food and Nutrition Data.

Jablonka, K. M., Ongari, D., Moosavi, S. M., & Smit, B. (2020). Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. Chemical Reviews, 120(16), 8066–8129. (https://dx.doi.org/10.1021/acs.chemrev.0c00004)

Jacobs, N., Brewer, S., Craigon, P. J., Frey, J., Gutierrez, A., Kanza, S., … Sacks, J. (2021). Considering the ethical implications of digital collaboration in the Food Sector. Patterns, 2(11). (https://dx.doi.org/10.1016/j.patter.2021.100335)

Jahin, M. A., Shovon, M. S. H., Shin, J., Ridoy, I. A., & Mridha, M. F. (2024). Big data—supply chain management framework for forecasting: Data preprocessing and machine learning techniques. *Archives of Computational Methods in Engineering, 31*(6), 3619-3645.(http://dx.doi.org/10.48550/arXiv.2307.12971)

Jadon, Aryan & Patil, Avinash & Jadon, Shruti. (2022). A Comprehensive Survey of Regression Based Loss Functions for Time Series Forecasting. (http://dx.doi.org/10.48550/arXiv.2211.02989)

Juul, F., Vaidean, G. and Parekh, N., 2021b. Ultra-processed foods and cardiovascular diseases: potential mechanisms of action. *Advances in Nutrition*, *12*(5), pp.1673-1680. (https://doi.org/10.1093/advances/nmab049)

Juul, F., Vaidean, G., Lin, Y., Deierlein, A.L. and Parekh, N., 2021a. Ultra-processed foods and incident cardiovascular disease in the Framingham Offspring Study. *Journal of the American College of Cardiology*, *77*(12), pp.1520-1531. (https://doi.org/10.1016/j.jacc.2021.01.047)

Kabir, M. M., Jim, J. R., & Istenes, Z. (2025). Terrain detection and segmentation for autonomous vehicle navigation: A state-of-the-art systematic review. *Information Fusion, 113*, 102644. (http://dx.doi.org/10.1016/j.inffus.2024.102644)

Kasapila, W., & Shaarani, S. M. (2011). Harmonisation of food labelling regulations in Southeast Asia: benefits, challenges and implications. Asia Pacific journal of clinical nutrition, 20(1), 1–8. (PMID: 21393103)

Kaviani, M., Almeida, J., & Verdi, F. L. (2024). Residual-based adaptive Huber loss (RAHL): Design of an improved Huber loss for CQI prediction in 5G networks. *arXiv preprint arXiv:2408.14718*. (https://doi.org/10.48550/arXiv.2408.14718)

Kenton, J.D.M.W.C. and Toutanova, L.K., 2019, June. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT* (Vol. 1, p. 2). (https://doi.org/10.18653/v1/N19-1423)

Keremidchiev, R. (2024). Optimizing machine learning models: Cost-effective feature selection for enhanced business performance. *Vanguard Scientific Instruments in Management, 20*(1), 223-237. (https://vsim-journal.info/index.php?journal=vsim&page=article&op=view&path[]=531)

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. (https://doi.org/10.48550/arXiv.1609.04836)

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I. & Goldstein, T.. (2023). A Watermark for Large Language Models. 202:17061-17084 (https://proceedings.mlr.press/v202/kirchenbauer23a.html)

Koido, M., Hon, C. C., Koyama, S., Kawaji, H., Murakawa, Y., Ishigaki, K., ... & Terao, C. (2023). Prediction of the cell-type-specific transcription of non-coding RNAs from genome sequences via machine learning. *Nature Biomedical Engineering, 7*(6), 830-844. (https://doi.org/10.1038/s41551-022-00961-8)

Kosma, C. (2023). *Towards Robust Deep Learning Methods for Time Series Data and their Applications* (Doctoral dissertation, Institut Polytechnique de Paris). (https://www.theses.fr/2023IPPAX140)

Lample, G., 2016. Neural architectures for named entity recognition. (https://doi.org/10.48550/arXiv.1603.01360)

Lane, M.M., Gamage, E., Travica, N., Dissanayaka, T., Ashtree, D.N., Gauci, S., Lotfaliany, M., O'neil, A., Jacka, F.N. and Marx, W., 2022. Ultra-processed food consumption and mental health: a systematic review and meta-analysis of observational studies. *Nutrients*, *14*(13), p.2568. (https://doi.org/10.3390/nu14132568)

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), pp.1234-1240. (https://doi.org/10.1093/bioinformatics/btz682)

Lei, T., Hu, Q., Hou, Z., & Lu, J. (2025). Enhancing real-world far-field speech with supervised adversarial training. *Applied Acoustics, 229*, 110407. (http://dx.doi.org/10.1016/j.apacoust.2024.110407)

Levy, R. B., Rauber, F., Chang, K., Louzada, M. L. da C., Monteiro, C. A., Millett, C., & Vamos, E. P. (2021). Ultra-processed food consumption and type 2 diabetes incidence: A prospective cohort study. Clinical Nutrition, 40(5), 3608–3614. (https://dx.doi.org/10.1016/j.clnu.2020.12.018)

Li, D., Yang, Y., Song, Y. Z., & Hospedales, T. (2018, April). Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1). (https://doi.org/10.1609/aaai.v32i1.11596)

Liu, M. and M'hiri, F., 2024, March. Beyond Traditional Teaching: Large Language Models as Simulated Teaching Assistants in Computer Science. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (pp. 743-749). (https://doi.org/10.1145/3626252.3630789)

Lu, Y., Stathopoulou, T., Vasiloglou, M.F., Pinault, L.F., Kiley, C., Spanakis, E.K. and Mougiakakou, S., 2020. goFOODTM: an artificial intelligence system for dietary assessment. *Sensors*, *20*(15), p.4283. (https://doi.org/10.3390/s20154283)

Lu, M., Rao, S., Yue, H., Han, J., & Wang, J. (2024). Recent advances in the application of machine learning to crystal behavior and crystallization process control. *Crystal Growth & Design, 24*(12), 5374–5396. (https://dx.doi.org/10.1021/acs.cgd.3c01251)

Ma, P., Li, A., Yu, N., Li, Y., Bahadur, R., Wang, Q., & Ahuja, J. K. (2021). Application of machine learning for estimating label nutrients using USDA Global Branded Food Products Database,(BFPD). *Journal of Food Composition and analysis, 100*, 103857. (https://doi.org/10.1016/j.jfca.2021.103857)

Ma, P., Zhang, Z., Li, Y., Yu, N., Sheng, J., McGinty, H. K., ... & Ahuja, J. K. (2022). Deep learning accurately predicts food categories and nutrients based on ingredient statements. *Food Chemistry, 391*, 133243. (https://doi.org/10.1016/j.foodchem.2022.133243)

Ma, P., Tsai, S., He, Y., Jia, X., Zhen, D., Yu, N., Wang, Q., Ahuja, J.K. and Wei, C.I., (2024). Large Language Models in Food Science: Innovations, Applications, and Future. *Trends in Food Science & Technology*, p.104488. (http://dx.doi.org/10.1016/j.tifs.2024.104488)

Marculescu, D., Stamoulis, D., & Cai, E. (2018, November). Hardware-aware machine learning: Modeling and optimization. In *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* (pp. 1-8). IEEE. (https://doi.org/10.48550/arXiv.1809.05476)

Martin, K., 2019. Designing Ethical Algorithms. MIS Quarterly Executive 18(2):129-142 (http://dx.doi.org/10.17705/2msqe.00012)

Mathew, A., Saldanha, A., & Babu, C. N. (2024). Audio–video syncing with lip movements using generative deep neural networks. *Multimedia Tools and Applications, 83*(35), 82019-82033. (https://doi.org/10.1007/s11042-024-18695-x)

Mavrogiorgos, K., Kiourtis, A., Mavrogiorgou, A., Menychtas, A., & Kyriazis, D. (2024). Bias in machine learning: A literature review. *Applied Sciences, 14*(19), 8860. (https://doi.org/10.3390/app14198860)

McElhinney, C. (2024) Development And Optimisation of Convolutional Neural Networks (Cnns) to Predict the Nutrition and Sustainability Scores of Foods from Crowd Sourced Images CCT College Dublin. (https://arc.cct.ie/msc_da/2/)

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR), 54*(6), 1–35. (https://doi.org/10.1145/3457607)

Meng, Q., Gu, J., & Liu, Y. H. (2024). GPD: Learning geometric primitive deformation for unseen object pose estimation. *IEEE Transactions on Automation Science and Engineering*. (https://doi.org/10.1109/TASE.2024.3514143)

Menghani, G. (2023). Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *ACM Computing Surveys, 55*(12), 1-37. (https://doi.org/10.1145/3578938)

Mente, A., O'Donnell, M. and Yusuf, S., 2021. Sodium intake and health: what should we recommend based on the current evidence?. *Nutrients*, *13*(9), p.3232. (https://doi.org/10.3390/nu13093232)

Minh, D., Wang, H.X., Li, Y.F. et al. Explainable artificial intelligence: a comprehensive review. Artif Intell Rev 55, 3503–3568 (2022). (https://doi.org/10.1007/s10462-021-10088-y)

Mienye, I. D., & Swart, T. G. (2024). A comprehensive review of deep learning: Architectures, recent advances, and applications. *Information, 15*(12), 755. (https://doi.org/10.3390/info15120755)

Minh, D., Wang, H.X., Li, Y.F. et al. Explainable artificial intelligence: a comprehensive review. Artif Intell Rev 55, 3503–3568 (2022). (https://doi.org/10.1007/s10462-021-10088-y)

Ming, Y., Yin, H., & Li, Y. (2022, June). On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(9), 10051-10059. (https://doi.org/10.1609/aaai.v36i9.21244)

Miraftabzadeh, S. A., Rad, P., Jamshidi, M., & Prevost, J. (2018, June). Customer review analytics using subjective loss function for conceptual-based learning. In *2018 13th Annual Conference on System of Systems Engineering (SoSE)* (pp. 211-218). IEEE. (https://doi.org/10.1109/SYSOSE.2018.8428702)

Miyazawa, T., Hiratsuka, Y., Toda, M., Hatakeyama, N., Ozawa, H., Abe, C., Cheng, T.Y., Matsushima, Y., Miyawaki, Y., Ashida, K. and Iimura, J., 2022. Artificial intelligence in food science and nutrition: a narrative review. *Nutrition Reviews*, *80*(12), pp.2288-2300.(https://doi.org/10.1093/nutrit/nuac033)

Mohebi, R., Chen, C., Ibrahim, N. E., McCarthy, C. P., Gaggin, H. K., Singer, D. E., Hyle, E. P., Wasfy, J. H., & Januzzi, J. L., Jr (2022). Cardiovascular Disease Projections in the United States Based on the 2020 Census Estimates. Journal of the American College of Cardiology, 80(6), 565–578. (https://doi.org/10.1016/j.jacc.2022.05.033)

Moradi, S., Entezari, M.H., Mohammadi, H., Jayedi, A., Lazaridi, A.V., Kermani, M.A.H. and Miraghajani, M., 2022. Ultra-processed food consumption and adult obesity risk: a systematic review and dose-response meta-analysis. *Critical reviews in food science and nutrition*, *63*(2), pp.249-260. (https://doi.org/10.1080/10408398.2021.1946005)

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate statistical machine learning methods for genomic prediction* (pp. 109–139). Cham: Springer International Publishing. (https://dx.doi.org/10.1007/978-3-030-89010-0_4)

Mostafa, W. M. M. (2024). *Time series classification of sport activities using neural networks* (Master's thesis). (https://elar.urfu.ru/bitstream/10995/140559/1/m_th_w.m.m.mostafa_2024.pdf

Mozaffarian D, Fleischhacker S, Andrés JR. Prioritizing Nutrition Security in the US. JAMA. 2021;325(16):1605–1606. (https://dx.doi.org/10.1001/jama.2021.1915

Naravane, T. (2024). Harnessing food composition data: Machine learning models to predict taste and health outcomes of food processing. *University of California, Davis.* (https://escholarship.org/uc/item/9d384027)

Naumenko, M., Hrashchenko, I., Nevmerzhytska, S., Tsalko, T., Krasniuk, S., & Kulynych, Y. (2024). Innovative technological modes of data mining and modeling for adaptive project management of food industry competitive enterprises in crisis conditions. *Project Management: Industry Specifics*. (https://er.knutd.edu.ua/bitstream/123456789/28140/1/SCOPUS_2024.pdf)

Nayak, R., & Waterson, P. (2019). Global food safety as a complex adaptive system: Key concepts and future prospects. *Trends in Food Science & Technology*, *91*, 409-425. (https://doi.org/10.1016/j.tifs.2019.07.040)

Neloy, A. A., & Turgeon, M. (2024). A comprehensive study of auto-encoders for anomaly detection: Efficiency and trade-offs. *Machine Learning with Applications*, 100572. (https://doi.org/10.1016/j.mlwa.2024.100572)

Ofodile, Onyeka & Toromade, Adekunle & Igwe, Abbey & Eyo-Udo, Nsisong & Olufemi-Phillips, Amarachi. (2024). Utilizing Predictive Analytics to Manage Food Supply and Demand in Adaptive Supply Chains. *Journal of Agricultural Economics and Management* (pending publication). (https://www.researchgate.net/publication/387312396_Utilizing_Predictive_Analytics_to_Manage_Food_Supply_and_Demand_in_Adaptive_Supply_Chains)

Pandove, D., Goel, S., & Rani, R. (2018). Systematic review of clustering high-dimensional and large datasets. *ACM Transactions on Knowledge Discovery from Data (TKDD), 12*(2), 1-68. (https://doi.org/10.1145/3132088)

Pagliai, G., Dinu, M., Madarena, M. P., Bonaccio, M., Iacoviello, L., & Sofi, F. (2021). Consumption of ultra-processed foods and health status: a systematic review and meta-analysis. British Journal of Nutrition, 125(3), 308–318. (https://dx.doi.org/10.1017/S0007114520002688)

Paulionis, L., 2008. The changing face of food and nutrition in Canada and the United States: opportunities and challenges for older adults. *Journal of Nutrition for the Elderly*, *27*(3-4), pp.277-295. (https://doi.org/10.1080/01639360802261979)

Pellegrini, C., Özsoy, E., Wintergerst, M., & Groh, G. (2021). Exploiting food embeddings for ingredient substitution. *HEALTHINF, 5*, 67–77. (https://doi.org/10.5220/0010202000670077)

Peters, C., Braschler, M. and Clough, P., 2012. *Multilingual information retrieval: From research to practice* (pp. I-XVII). Heidelberg: Springer. (https://doi.org/10.1007/978-3-642-23008-0)

Pillay, M. T., Minakawa, N., Kim, Y., Kgalane, N., Ratnam, J. V., Behera, S. K., ... & Sweijd, N. (2023). Utilizing a novel high-resolution malaria dataset for climate-informed predictions with a deep learning transformer model. *Scientific Reports, 13*(1), 23091. (http://dx.doi.org/10.1038/s41598-023-50176-3)

Piles, M., Bergsma, R., Gianola, D., Gilbert, H., & Tusell, L. (2021). Feature selection stability and accuracy of prediction models for genomic prediction of residual feed intake in pigs using

machine learning. *Frontiers in Genetics, 12*, 611506. (https://doi.org/10.3389/fgene.2021.611506)

Poldrack, R. A., Huckins, G., & Varoquaux, G. (2020). Establishment of best practices for evidence for prediction: A review. *JAMA Psychiatry, 77*(5), 534–540. (https://doi.org/10.1001/jamapsychiatry.2019.3671)

Popkin, B.M. and Ng, S.W., 2022. The nutrition transition to a stage of high obesity and noncommunicable disease prevalence dominated by ultra-processed foods is not inevitable. *Obesity Reviews*, *23*(1), p.e13366. (https://doi.org/10.1111/obr.13366)

Popovski, G., Seljak, B.K. and Eftimov, T., 2020. A survey of named-entity recognition methods for food information extraction. *IEEE Access*, *8*, pp.31586-31594. (http://dx.doi.org/10.1109/ACCESS.2020.2973502)

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. (https://doi.org/10.48550/arXiv.2103.00020)

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1-67. (https://doi.org/10.48550/arXiv.1910.10683)

Rane, N. L., Mallick, S. K., Kaya, Ö., & Rane, J. (2024). *Applied machine learning and deep learning: Architectures and techniques.* Deep Science Publishing. (https://doi.org/10.70593/978-81-981271-4-3)

Rastegar, S., Doughty, H., & Snoek, C. (2023). Learn to categorize or categorize to learn? Self-coding for generalized category discovery. *Advances in Neural Information Processing Systems, 36*, 72794-72818. (https://doi.org/10.48550/arXiv.2310.19776)

Rauber, F., Chang, K., Vamos, E.P., da Costa Louzada, M.L., Monteiro, C.A., Millett, C. and Levy, R.B., 2021. Ultra-processed food consumption and risk of obesity: a prospective cohort study of UK Biobank. *European journal of nutrition*, *60*, pp.2169-2180. (https://doi.org/10.1007/s00394-020-02367-1)

Rebelo, J. E. L. (2024). Enhancing interpretability of neural networks in food recommendation systems (Master's thesis, Instituto Politecnico de Viseu, Portugal). (http://hdl.handle.net/10400.19/8454)

Rodríguez, A.L.B., Amarilla, N.J.D., Rodríguez, M.M.T., Martínez, B.E.N. and Meza-Miranda, E.R., 2022. Processed and ultra-processed foods consumption in adults and its relationship with quality of life and quality of sleep. *Revista de Nutrição*, *35*, p.e220173. (https://doi.org/10.1590/1678-9865202235e220173)

Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., & Schmidt, L. (2019). A meta-analysis of overfitting in machine learning. *Advances in Neural Information Processing Systems, 32*. (https://proceedings.neurips.cc/paper_files/paper/2019/file/ee39e503b6bedf0c98c388b7e8589aca-Paper.pdf)

Ropodi, A. I., Panagou, E. Z., & Nychas, G. J. (2016). Data mining derived from food analyses using non-invasive/non-destructive analytical techniques; determination of food authenticity, quality & safety in tandem with computer science disciplines. *Trends in Food Science & Technology, 50*, 11-25. (http://dx.doi.org/10.1016/j.tifs.2016.01.011)

Roy, R., Marakkar, S., Vayalil, M. P., Shahanaz, A., Anil, A. P., Kunnathpeedikayil, S., ... & Yadav, K. K. (2022). Drug-food interactions in the era of molecular big data, machine intelligence, and personalized health. *Recent Advances in Food Nutrition & Agriculture, 13(1),* 27-50. (http://dx.doi.org/10.2174/2212798412666220620104809)

Salim, N. O., & Mohammed, A. K. (2024). Comparative Analysis of Classical Machine Learning and Deep Learning Methods for Fruit Image Recognition and Classification. *Traitement du Signal, 41(3). (https://doi.org/10.18280/ts.410322)*

Salam, M. A., Azar, A. T., Elgendy, M. S., & Fouad, K. M. (2021). The effect of different dimensionality reduction techniques on machine learning overfitting problem. *International Journal of Advanced Computer Science and Applications, 12*(4), 641-655. (https://dx.doi.org/10.14569/IJACSA.2021.0120480)

Santilal, U., 2020. *Natural Language Processing: NLP & its History (http://dx.doi.org/10.1007/978-981-15-9712-1_31)*

Santos, C. F. G. D., & Papa, J. P. (2022). Avoiding overfitting: A survey on regularization methods for convolutional neural networks. *ACM Computing Surveys (CSUR), 54*(10s), 1-25. (https://doi.org/10.1145/3510413)

Shah, M., & Sureja, N. (2025). A comprehensive review of bias in deep learning models: Methods, impacts, and future directions. *Archives of Computational Methods in Engineering, 32*(1), 255–267. (https://doi.org/10.1007/s11831-024-10134-2)

Shi, Y., Wang, X., Chen, S., Zhao, Y., Wang, Y., Sheng, X., ... & Xing, K. (2025). Identification of key genes affecting intramuscular fat deposition in pigs using machine learning models. *Frontiers in Genetics, 15*, 1503148. (https://doi.org/10.3389/fgene.2024.1503148)

Shoaib, M. R., Emara, H. M., & Zhao, J. (2023). Revolutionizing global food security: Empowering resilience through integrated AI foundation models and data-driven solutions. *arXiv preprint arXiv:2310.20301. (https://doi.org/10.48550/arXiv.2310.20301)*

Siddique, A., Cook, K., Holt, Y., Panda, S. S., Mahapatra, A. K., Morgan, E. R., ... & Terrill, T. H. (2024). From plants to pixels: The role of artificial intelligence in identifying *Sericea Lespedeza* in field-based studies. *Agronomy, 14*(5), 992. (http://dx.doi.org/10.20944/preprints202404.1002.v1)

Singhal, P., Walambe, R., Ramanna, S., & Kotecha, K. (2023). Domain adaptation: Challenges, methods, datasets, and applications. *IEEE Access, 11*, 6973-7020. (https://ieeexplore.ieee.org/iel7/6287639/6514899/10017290.pdf)

Smith, S. L., Kindermans, P. J., Ying, C., & Le, Q. V. (2017). Don't decay the learning rate, increase the batch size. (https://doi.org/10.48550/arXiv.1711.00489)

Suddul, G., & Seguin, J. F. L. (2023). A comparative study of deep learning methods for food classification with images. *Food and Humanity, 1*, 800-808. (https://doi.org/10.1016/j.foohum.2023.07.018)

Suksatan, W., Moradi, S., Naeini, F., Bagheri, R., Mohammadi, H., Talebi, S., Mehrabani, S., Hojjati Kermani, M.A. and Suzuki, K., 2021. Ultra-processed food consumption and adult mortality risk: a systematic review and dose–response meta-analysis of 207,291 participants. *Nutrients*, *14*(1), p.174. (http://dx.doi.org/10.3390/nu14010174)

Tai, R.H., Bentley, L.R., Xia, X., Sitt, J.M., Fankhauser, S.C., Chicas-Mosier, A.M. and Monteith, B.G., 2024. An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, *23*, p.16094069241231168. (http://dx.doi.org/10.1177/16094069241231168)

Taha, K. (2024). Employing machine learning techniques to detect protein function: A survey, experimental, and empirical evaluations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics. (http://dx.doi.org/10.1109/TCBB.2024.3427381)*

Tao, D., Yang, P., & Feng, H. (2020). Utilization of text mining as a big data analysis tool for food science and nutrition. *Comprehensive Reviews in Food Science and Food Safety, 19*(2), 875–894. (https://doi.org/10.1111/1541-4337.12540)

Theng, D., & Bhoyar, K. K. (2024). Feature selection techniques for machine learning: A survey of more than two decades of research. *Knowledge and Information Systems, 66*(3), 1575-1637. (http://dx.doi.org/10.1007/s10115-023-02010-5)

Thakkar, A., & Lohiya, R. (2022). A survey on intrusion detection system: Feature selection, model, performance measures, application perspective, challenges, and future research directions. *Artificial Intelligence Review, 55*(1), 453-563. (https://link.springer.com/article/10.1007/s10462-021-10037-9)

Temple, N.J., 2020. Front-of-package food labels: A narrative review. *Appetite*, *144*, p.104485. (https://doi.org/10.1016/j.appet.2019.104485)

Teng, M.L., Ng, C.H., Huang, D.Q., Chan, K.E., Tan, D.J., Lim, W.H., Yang, J.D., Tan, E. and Muthiah, M.D., 2023. Global incidence and prevalence of nonalcoholic fatty liver disease. *Clinical and molecular hepatology*, *29*(Suppl), p.S32. (https://doi.org/10.3350/cmh.2022.0365)

Tiozon, R. J. N., Sreenivasulu, N., Alseekh, S., Sartagoda, K. J. D., Usadel, B., & Fernie, A. R. (2023). Metabolomics and machine learning technique revealed that germination enhances the multi-nutritional properties of pigmented rice. *Communications Biology, 6*(1), 1000. (http://dx.doi.org/10.1038/s42003-023-05379-9)

Tristan Asensi, M., Napoletano, A., Sofi, F. and Dinu, M., 2023. Low-grade inflammation and ultra-processed foods consumption: a review. *Nutrients*, *15*(6), p.1546. (https://doi.org/10.3390/nu15061546)

Tsai, H., Riesa, J., Johnson, M., Arivazhagan, N., Li, X. and Archer, A., 2019. Small and practical BERT models for sequence labelling. (https://doi.org/10.48550/arXiv.1909.00100)

Tsoy, A., Liu, Z., Zhang, H., Zhou, M., Yang, W., Geng, H., ... & Geng, Z. (2024). Image-free single-pixel keypoint detection for privacy-preserving human pose estimation. *Optics Letters, 49*(3), 546-549. (https://doi.org/10.1364/OL.514213)

Turing, A.M., 1936. On computable numbers, with an application to the Entscheidungsproblem. *J. of Math*, *58*(345-363), p.5. (https://www.cs.virginia.edu/~robins/Turing_Paper_1936.pdf)

Umar, I. H., Salga, M. S., Lin, H., Hassan, J. I., Ahmad, A., Ibrahim, A. S., & Jechira, B. K. (2025). Performance characterisation of machine learning models for geotechnical axial pile load capacity estimation: An enhanced GPR-based approach. *Geomechanics and Geoengineering*, 1-42. (https://www.tandfonline.com/doi/full/10.1080/17486025.2025.2468645)

Van den Wijngaart, A.W., 2002. Nutrition labelling: purpose, scientific issues and challenges. *Asia Pacific journal of clinical nutrition*, *11*(2), pp.S68-S71. (https://doi.org/10.1046/j.1440-6047.2002.00001.x)

Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research, 144*, 93–106. (http://dx.doi.org/10.1016/j.jbusres.2022.01.076)

Van Leeuwen, B., & Nutzel, M. (2024). Detecting drug transfers via the drop-off method: A supervised model approach using AIS data. *Machine Learning with Applications, 18*, 100590. (https://doi.org/10.1016/j.mlwa.2024.100590)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30.* (https://doi.org/10.48550/arXiv.1706.03762)

Vijayakumar, G., & Bharathi, R. K. (2024). OptiFeat: Enhancing feature selection, a hybrid approach combining subject matter expertise and recursive feature elimination method. *Discover Computing, 27*(1), 1-13. (https://doi.org/10.1007/s10791-024-09483-0)

Wang, X., Zhang, J., Xun, L., Wang, J., Wu, Z., Henchiri, M., ... & Yu, X. (2022). Evaluating the effectiveness of machine learning and deep learning models combined time-series satellite data for multiple crop types classification over a large-scale region. *Remote Sensing, 14*(10), 2341. (https://doi.org/10.3390/rs14102341)

Wang, H., Ullah, Z., Gazit, E., Brozgol, M., Tan, T., Hausdorff, J. M., ... & Ponger, P. (2024). Step Width Estimation in Individuals With and Without Neurodegenerative Disease Via a Novel Data-Augmentation Deep Learning Model and Minimal Wearable Inertial Sensors. *IEEE Journal of Biomedical and Health Informatics*. (https://ieeexplore.ieee.org/document/10697468/)

Wei, H., Gao, M., Zhou, A., Chen, F., Qu, W., Wang, C. and Lu, M., 2019. Named entity recognition from biomedical texts using a fusion attention-based BiLSTM-CRF. *IEEE Access*, *7*, pp.73627-73636. (http://dx.doi.org/10.1109/ACCESS.2019.2920734)

Weng, W. H. (2020). Machine learning for clinical predictive analytics. In *Leveraging Data Science for Global Health* (pp. 199-217). (https://doi.org/10.1007/978-3-030-47994-7_12)

Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2015). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation, 20*(4), 606-626. (https://doi.org/10.1109/TEVC.2015.2504420)

Yadav, V. and Bethard, S., 2019. A survey on recent advances in named entity recognition from deep learning models. (https://doi.org/10.48550/arXiv.1910.11470)

Yan, Y., Zheng, P., & Wang, Y. (2024). Enhancing large language model capabilities for rumor detection with knowledge-powered prompting. *Engineering Applications of Artificial Intelligence, 133*, 108259. (https://doi.org/10.1016/j.engappai.2024.108259)

Yang, Y., Lou, H., Wang, Z., & Wu, J. (2024). Pinball-Huber boosted extreme learning machine regression: A multi-objective approach to accurate power load forecasting. *Applied Intelligence, 54*(17), 8745-8760. (https://doi.org/10.1007/s10489-024-05651-3)

Yaqoob, A., & Muntean, G. M. (2024). FReD-ViQ: Fuzzy Reinforcement Learning Driven Adaptive Streaming Solution for Improved Video Quality of Experience. *IEEE Transactions on Network and Service Management*. (http://dx.doi.org/10.1109/TNSM.2024.3450014)

Ye, W., Zheng, G., Cao, X., Ma, Y., & Zhang, A. (2024). Spurious correlations in machine learning: A survey. (https://doi.org/10.48550/arXiv.2402.12715)

Yu, J., Li, Y., Liu, Z., & Yang, Q. (2025). Adaptive graph neural network protection algorithm based on differential privacy. *Journal of Systems and Software*, 112386. (https://doi.org/10.1016/j.jss.2025.112386)

Zambrano Chaves, J.M., Wang, E., Tu, T., Dhaval Vaishnav, E., Lee, B., Mahdavi, S.S., Semturs, C., Fleet, D., Natarajan, V. and Azizi, S., 2024. Tx-LLM: A Large Language Model for Therapeutics. (https://doi.org/10.48550/arXiv.2406.06316)

Zhang, A., Ballas, N., & Pineau, J. (2018). A dissection of overfitting and generalization in continuous reinforcement learning. (https://doi.org/10.48550/arXiv.1806.07937)

Zhang, Y., Bao, X., Zhu, Y., Dai, Z., Shen, Q., & Xue, Y. (2024). Advances in machine learning screening of food bioactive compounds. *Trends in Food Science & Technology*, 104578. (https://doi.org/10.1016/j.tifs.2024.104578)

Zerouali, B., Bailek, N., Tariq, A., Kuriqi, A., Guermoui, M., Alharbi, A. H., ... & El-Kenawy, E. S. M. (2024). Enhancing deep learning-based slope stability classification using a novel metaheuristic optimization algorithm for feature selection. *Scientific Reports, 14*(1), 21812. (http://dx.doi.org/10.1038/s41598-024-72588-5)

Zhu, J. J., Yang, M., & Ren, Z. J. (2023). Machine learning in environmental research: Common pitfalls and best practices. *Environmental Science & Technology, 57*(46), 17671-17689. (https://doi.org/10.1021/acs.est.3c00026)

Zhou, P., Min, W., Fu, C., Jin, Y., Huang, M., Li, X., Mei, S. and Jiang, S., 2024. FoodSky: A Food-oriented Large Language Model that Passes the Chef and Dietetic Examination. (https://doi.org/10.48550/arXiv.2406.10261)