Healthy/ Non-Healthy Food Classification via Nutrient Content Image Recognition in Packed food

Healthy Food Prediction leveraging OCR and Machine Learning

Diya Srivastava
Master's in Artificial Intelligence
National College of Ireland
Dublin, Ireland
x23177608@student.ncirl.ie

***Abstract*—In the era of abundant data, leveraging Machine Learning to acquire extensive insights from raw information has become a transformative aspect of Artificial Intelligence paradigm. Despite the availability of Nutrient information on packed food items, we tend to ignore the health aspect of it due to lack of knowledge, time and expertise to interpret this data efficiently. Hence, to foster this problem with a comprehensive solution this research proposes a Machine learning approach to classify Food items as Healthy or Not Healthy through Image Recognition of Nutrition-Table on packed food. For dataset, I have managed to collect 580 images of different food items present in Dublin grocery stores. This study utilizes Optical Character Recognition (OCR) to extract textual data i.e. Macro-Nutrients and their quantities. And leveraged Random Forest, Logistic Regression, SVM, Gradient Boosting and Decision Tree Classifier for Healthy, Non-Healthy prediction. These prediction results are evaluated by accuracy, precision and f1-score Metrics.To achieve 95% accuracy by Random Forest Classifier, 94% accuracy by Logistic Regression followed by rest and finally 97% accuracy attained by Gradient Boosting Classifier.**

Keywords: Nutri-Score

I. Introduction

Machine Learning (ML) and the application of science and algorithms, that understand data and take out intriguing insights and patterns from them is one of the most useful and game changing aspect of the Artificial Intelligence paradigm. We are living in the age where every information is potential data and can be utilised to extract information and formulate it into an idea worth of billion dollar. Leveraging algorithms of Supervised, Unsupervised and Reinforcement learning, one can develop models to classify, predict, find hidden patterns and do a lot more. This research is based on the exact idea of extracting useful information from raw data and make use of it for betterment.

Do we really understand or scrutinize the food that we buy and consume? There's quite a lot of intriguing ingredients and macro-nutrient information mentioned on the packed food that we buy, that most of us never read or have the time spending to understand the Nutrient Content Mathematics. But, that Informative insight can tell us a lot about what we're consuming. And our body is made up of what we serve it, what's the fuel we choose for ourselves. Leading us to the reliance and significance of Nutrients in food and Human lives. If we scrutinize the statistics, there would be more than half of the population diagnosed with a lifestyle disorder oriented by Nutrient Insufficiency, Increased Fat percentage and improper management of one's weight. To support the fact, as stated by National Clinical Programs of Health Safety and Environment (HSE) [6] , Ireland has the highest levels of Obesity in Europe, over 60% of adults and 1 in five children are suffering from obesity, which is associated with other chronic diseases like diabetes, hypothyroidism, etc. According to the National Diabetes Registry survey [2] of 2022, over 3.1 million people of the population of 5.2 million in Ireland were diabetic, that is approximately 60% of the population, and these are only few of the health disorders that are prevailed by Insufficient Dietary needs. The Department of Health and Children [4], estimates that there are currently 200,000 people in Ireland affected by eating disorders. Nutrients in the right quantity play a vital role in determining the health of an individual. Nutrients like "Carbohydrates", "Fats" and "Proteins" are crucial for energy balance and weight management, whereas "Fibre", vitamins and antioxidants are important for prevention of Chronic disease, on the other hand vitamins and calcium are important for healthy bones and Immunity.

This quite serves the incentive to develop a comprehensive solution such that one gets to know the Health benefits of packed food items just by taking a picture of the nutrition table or nutrient content of them and getting to know the Healthy or Non- Healthy aspect of the food item. Not everyone can afford a nutritionist or a dietary plan that would achieve their motive of gaining or loosing weight. This could be an extensive solution to classify food items as healthy/Non healthy in real time through an image of the packet.

Gathering Nutrient Insights, drawing analysis and providing Factual information is not a new approach, instead its a continuum of several Nutrition Oriented Applications. A few works like, [9] Open Food Facts have been implemented on large scale, where anyone can add Nutrient Table im-

age to the database and access other Food item's Nutrient content and insights like Nutri-Score, Ingredients, Additives and Food Processing by scanning the Bar-code of Food items on packaging. But, only for items that are present in Open-Food-Facts database. Working on this limitation, although for classification of Healthy Food Items this research aims to scan an image of Nutrient-Table, get Textual data through Optical Character Recognition run Machine Learning Classifier models like Random Forest, Logistic Regression, Gradient Boosting etc. on this data and classify as Healthy or Non-Healthy.

II. Literature Review

Research in this domain is not a completely new concept, but is a continuum of several diverse studies.There are a few niche research works which have leveraged ML and Deep Learning (DL) algorithms to understand the Nutrients, ingredients and correlation amongst these quantities to yield or predict useful information about the food we consume. Where, [9] is one of the most extensively comprehensive solution, providing every detail of the food item by a simple scan of bar-code from packed food. Open-Food-Facts is a non-profit association of volunteers with over 25.000+ contributors. It is an established database for everyone and by everyone. The potential limitation of this study is its scalability and lack of availability due to very few products registered with the association and hence a detailed information about only these products. While, this study proposes a real-time Healthy Food prediction system that does not rely on a database. Leveraging OCR technique, this prediction system will extract useful information from packed food item, run analysis on this data and yield useful information.

The first task is Optical Character Recognition for which there are numerous libraries and API's available like Tesseract, EasyOCR, PaddleOCR and Google Vision, etc. The first task is to extract Text from Images, on careful evaluation of all these OCR tools, I have leveraged Google Vision API because of its accurate Text Recognition, that will be discussed in later sections. There has been several research on Nutrient Extraction from the food packet a few of these are, like JULIA REIBRING [14] in her Master's Thesis worked on "Photo OCR on American Nutrition table" and achieved a Recall and Precision of 0.72 and 0.82 and an accuracy of 89%. She had used the Convolutional Neural Network (CNN) for Image-to-text conversion and Text detection ans Extraction from the image, she used CNN with the idea of Contour detection as she worked with Nutrition table that were in table format so, text recognition becomes easier by detecting these rectangles of table. Although these Neural Networks like CNN are good for image recognition, detection tasks and are not suitable for complex Text Reading tasks as required in this study. Another work in the same field is done by (Yaksh et al.) [15] in the work "Automated Nutrition Table Extraction and Ingredient Recognition" where they leveraged PaddleOCR and TensorFlow Object Detection API for text extraction from the images and provided Allergic Information, Additives and Nutrient description for the scanned item. Another interesting work done by (A. Parkavi, et al.) [12] in their work "Android application for food label recognition to ensure safe food consumption based on user allergen information leveraging OCR", where they have developed an android application that recommends the user to consume the food or not based on user allergens through reading the food packaging and checks for allergens, leveraging Google vision API provided by firebase API and achieved an accuracy of 91% for OCR and 99% for allergen detection. A great contribution by (Marwa Ahmed, et al.) [1] in their work "Development of the Food Label Information Program: A Comprehensive Canadian Branded Food Composition Database" introduced Food Label Information Program (FLIP),have utilized big data approach to evaluate the Canadian food supply and presents the latest methods used in the development of this database, they had over 70,000¿ dataset since they were performing OCR along with web scrapping, and achieved good results in the same. Although Web scraping is a good idea, it is again a complex task to acquire data from [9] because the Nutrient-Table and related content is not specifically in English, and non-uniformity in an inconsistent dataset is also a barrier which one has to overcome only by revising acquired data.

In a survey paper, a research done by (S. Kayalvizhi et al.) [7] stated as "Product Constituents and FDA regulations for Consumer Safety Using OCR", is a work leveraging OCR to entail details of a product to identify if its fit for consumption since a few companies hide some relevant information about allergens. So, they have used OCR to extract text, and Semantic search and Deep Neural Network to extract a user profile after extracting characteristics from the analysis of a component, also Natural language processing to comprehend human language and emotions. They concluded the most hidden ingredient unfit for consumption in these foods are Phthalates, BPA, PBDE and Lead. Hence, serves as an inspiration to scrutinize what one can achieve in the same field. Another research that is not directly related to my work but can yet be useful "OCR-and-ML-Classification-of-Shopping-Receipts" by (Benedikt et al.) [8] deals with the processing of shopping receipts leveraging OCR to determine 5-digit Classification of Individual Consumption by Purpose (COICOP). And, they've recieved tremendously different results using deep learning for image scanning, then performed image preprocessing, followed by OCR, NLP and ML classification to achieve useful manipulation of receipt images put to use. In the research they have made use of Levenshtein distance to calculate accuracy of OCR results which came out to be 99% and 89% for few texts. And, a final AI model for the classification task with an accuracy of 80%. finally, "Intro to OCR- Tesseract, Opencv.." this book by PyImageSearch [13] is a great inspiration to work with optical character recognition and has a lot of applications and insights to various libraries of OCR. For Classification and prediction tasks, there have been various commendable researches as well in the Food and its Nutrition domain. In the work 'Deep Learning Based Nutrient-Driven Categorization of Packaged Food Sauces: Enhancing Consumer Awareness through Rule-

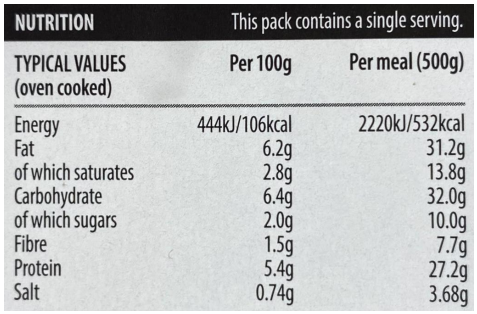
Based
Classification' [16] by (T Kusuma et al., 2024) have leveraged Artificial Neural Networks (ANN) to classify food nutrients composition of packaged food sauces and classifying the food into their corresponding nutrients composition such as fats, carbohydrates, proteins and more. For database they have utilized open-source food sauce nutrients database with nutritional labels, and ingredient lists. Also, leveraged SVM for Nutrition score labelling, to finally achieve accuracy rate of 85% from SVM and 82% from ANN classification. Takeaway from this study was the efficient performance of ML classification model SVM over Neural Networks ANN, motivating to work with ML models for classification and prediction tasks when small dataset is available.

There have been several research and studies leveraging Deep Learning methods in Food processing domain , like in the study 'Deep learning and machine vision for food processing' by (Lili Zhu, et al., 2021) they are working with image processing and machine vision to predict the quality of food through Image recognition and several characteristics of the same like Nutrients, ingredients etc., in this survey paper the researchers have worked on tasks such as food grading, detecting locations of defective spots or foreign objects, and removing impurities.Finally, they have covered a lot of approaches to identify Food Quality majoring Image processing and Recognition leveraging CNN, ANN, assigning these food items into categories by KNN and SVM and carrying out predictions using SVM and logistic regressions. All of which helped me to formulate an idea of where to focus in terms of Methodology, and I concluded with ML models. Finally, a research namely -'Automatic Estimation of the Nutritional Composition of Foods as part of the Glucose ML Type 1 Diabetes Self-Management System' by (Fotis K. et al, 2019) leverages a computer-vision-based approach that outlined combining image processing and machine learning to plate detection, food segmentation, food recognition and volume estimation of a plate's content for short-term predictive analytics of the glucose trajectory. For database, the study has utilized the GlucoseML food image database which relies on Greek food composition images of seafoods, milk foods, grains and much more. Although the project is still in design phase, Deep Learning and ML models are applied in feature (key points) extraction,food classification and 3D reconstruction. The method implementation is based on OpenCV, VLFeat, and ScikitLearn libraries

III. Dataset Preparation and Statistics

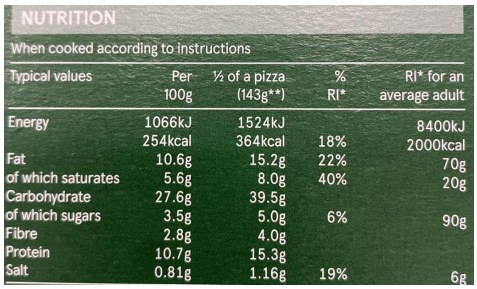
A. *Data Source*

The dataset utilized in this study is images of the Nutrition table of packed food items. These images are of the back of any packed food item with the Nutrition Content information in it. I have captured images manually from the grocery stores available in Dublin. A total of 580 images in HEIC format, which currently account for 150-500 KB have been collected. These are clear, colored pictures from recent packed foods available at supermarkets or grocery stores like Tesco, Aldi, Dunnes and Lidl etc. The device used for data collection is Iphone 11.



NUTRITION	This pack contains a single serving.	
TYPICAL VALUES (oven cooked)	Per 100g	Per meal (500g)
Energy	444kJ/106kcal	2220kJ/532kcal
Fat	6.2g	31.2g
of which saturates	2.8g	13.8g
Carbohydrate	6.4g	32.0g
of which sugars	2.0g	10.0g
Fibre	1.5g	7.7g
Protein	5.4g	27.2g
Salt	0.74g	3.68g

Fig. 1. aldi cooked cabbage carrot Meal

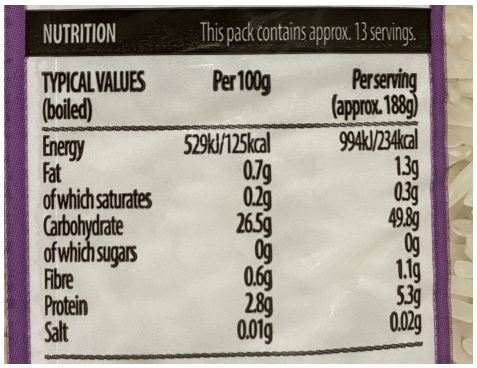


NUTRITION

When cooked according to instructions

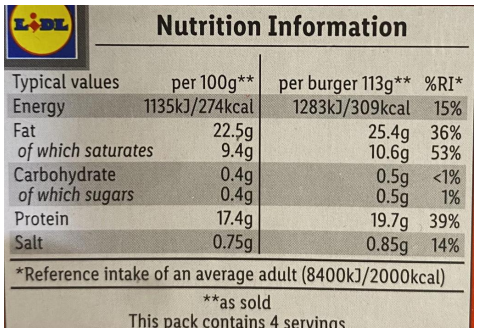
Typical values	Per 100g	½ of a pizza (143g**)	% RI*	RI* for an average adult
Energy	1066kJ 254kcal	1524kJ 364kcal	18%	8400kJ 2000kcal
Fat	10.6g	15.2g	22%	70g
of which saturates	5.6g	8.0g	40%	20g
Carbohydrate	27.6g	39.5g		
of which sugars	3.5g	5.0g	6%	90g
Fibre	2.8g	4.0g		
Protein	10.7g	15.3g		
Salt	0.81g	1.16g	19%	6g

Fig. 2. Tesco Stonebaked Piri-piri Chicek Pizza



NUTRITION	This pack contains approx. 13 servings.	
TYPICAL VALUES (boiled)	Per 100g	Per serving (approx. 188g)
Energy	529kJ/125kcal	994kJ/234kcal
Fat	0.7g	1.3g
of which saturates	0.2g	0.3g
Carbohydrate	26.5g	49.8g
of which sugars	0g	0g
Fibre	0.6g	1.1g
Protein	2.8g	5.3g
Salt	0.01g	0.02g

Fig. 3. Aldi Basmati Rice packet



Nutrition Information

Typical values	per 100g**	per burger 113g**	%RI*
Energy	1135kJ/274kcal	1283kJ/309kcal	15%
Fat	22.5g	25.4g	36%
of which saturates	9.4g	10.6g	53%
Carbohydrate	0.4g	0.5g	<1%
of which sugars	0.4g	0.5g	1%
Protein	17.4g	19.7g	39%
Salt	0.75g	0.85g	14%

*Reference intake of an average adult (8400kJ/2000kcal)

**as sold

This pack contains 4 servings

Fig. 4. Nutrition Table Sample 1 of Chocuer Milk Chocolate

B. *Data Preparation*

- **Image Conversion:** The original format of the images that I managed to capture is HEIC format, which is unsupported by many pandas functions. So, I changes it to jpg/jpeg format using the pillow-heif converter.
- **Crop Images:** The images were not cropped initially when taken, so they were cropped to keep the portion needed in the frame, that contains the Nutrient content.
- **Image Resize:** Finally, for OCR text recognition, to make optimal use of the API or OCR library the image size should be in range to accommodate as much as 580 images. For this purpose using pillow library only into 1024x1024 pixels.

IV. Methodology and Techniques

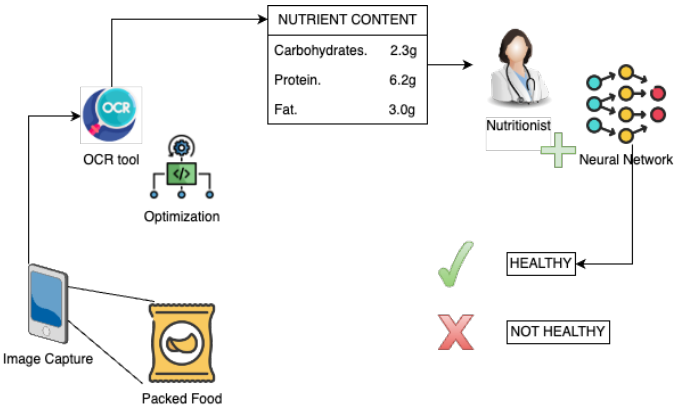


Fig. 5. OCR Methodology

A. *Optical Character Recognition*

*1) **Objective**:* Extract Text from Images. Here, the motive and first task for the data collection and Classification model is to extract Nutrient Content from the package of edibles. Where, first I worked on Reading the text accurately leveraging the best performing library amongst the one's listed below from only the Nutrition Table. And finally, after this computer vision task of Nutrient Recognition from the image, I leveraged ML models for classification of the food item into Healthy or Non-healthy food based on its Nutrient Content.

*2) **Traditional OCR Methods**:*

- Tesseract: "Tesseract OCR" or Pytesseract is an open source OCR engines that uses a bunch of Machine Learning algorithms to extract text from an image, supporting over 100+ languages [17]. I have worked with this library with my data as used by (Louisa, et al.) [8] as they achieved an average accuracy of 90% for OCR results, but it didn't give good results as it performed very bad in reading complex and detailed text from images. So, I moved forward with another OCR library.
- EasyOCR: [3] This is another Text reading tool from images that supports 80+ languages and is integrated into HuggingFace, which is relatively new than pytesseract and is suspected to be easier to implement giving better results. This gave relatively better results than tesseract, but was not 100% accurate. As results, this did read complex text better than Tesseract but failed to read quantities with utmost precision. Since, that was a major requirement of the project and couldn't be neglected I moved to another better option available for complex and detailed text tracking and reading.
- PaddleOCR: This is also a text reading tool that was launched in 2020 and has been updated regularly with its recent version in 2023, it allegedly gives great accuracy and supports 80 multilingual [11]. And working with this gave the most accurate results as achieved by (Shah, et al.) [15]. Although this is one of the best Detailed and extremely small Text Reader for OCR tasks, it although read 90% of the text accurately but failed to read the quantities precisely which was the drawback and couldn't be neglected.
- Google Vision: Google Vision API is one of the best resorts to capture text in image through its OCR feature [5]. Although has provided accuracy of 99% and 91% for different texts in the research by (Parkavi, et al.) [12]. Leveraging this as done by [9], this gave a 100% accurate results even for smallest most detailed text reading tasks especially reading quantities of respective macro-nutrients with utmost accuracy.
- Open CV: [10] This is one of the infamous libraries of programming functions for real time computer vision. I tried using the OpenCV Open-source library for image reading and display, image preprocessing, reading and detecting text contours and changing from RGB to BW since OCR did not benefit from the color of the image. But, none of the OCR libraries yielded a 100% accuracy this wasn't of much significance and for Google Vision I didn't have to use OpenCV.

*3) **Steps**:* The steps to achieve textual data from image utsing libraries like Tesseract or EasyOCR involves:

- Data preprocessing: Converting Images to grey scale as colour isn't an important factor here. Resizing the images to serve consistent and light images to the model.
- Character Segmentation: Segregating the image into individual character.
- Contour Detection: This is one of the most significant step to recognise the boundary of text in the image so that model can read it accurately.
- Text read: Finally read the desired text from the contours.

B. *Healthy or Non-Healthy Classification*

*1) **Objective**:* This is the step after data collection is done, when we have 580 images of the packed food nutrition table. The objective here is Classification into Healthy and Non-Healthy food items based on the Nutrient Content in them. This can be achieved once we have the textual data extracted from the image that we have done above.

*2) **Data Augmentation**:*

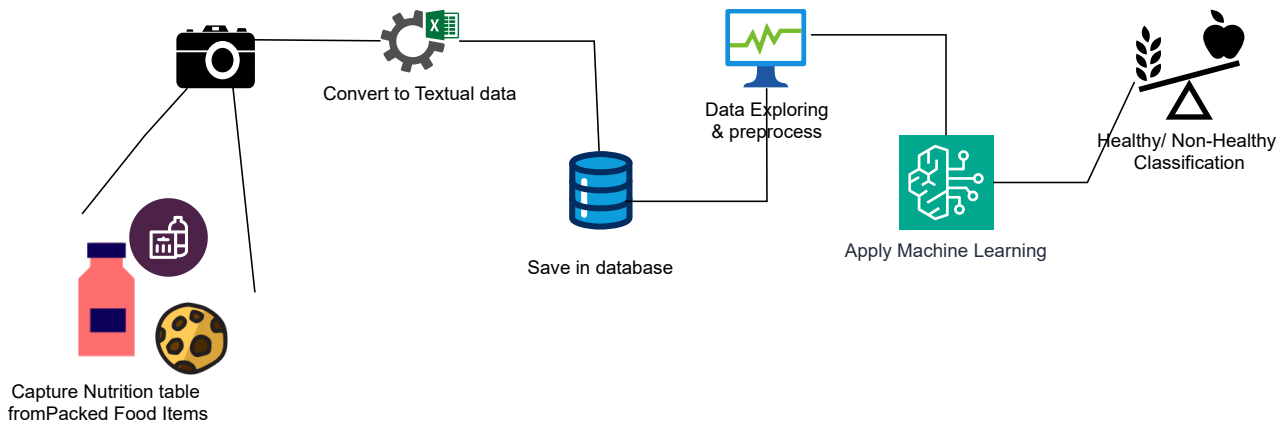


Fig. 6. Working of OCR+ ML model

- **Fetch Data:** After Image collection and OCR of relevant text, its important to fetch available Nutrients and their quantities from the piece of text. therefore, I leveraged regular expression for each nutrient and its corresponding quantities in different measuring units for 100gms of food product.
- **Textual Conversion:** After Regular Expression matches, its important to save this information in relevant and structured format. Hence, I saved it in a dataframe with respective food Product name that I labelled it with. This step involves cleaning data, removing irrelevant symbols and measuring units like gm or l and storing numerical values.
- **Handle Missing Values:** Since the data is really small and the only information I had to predict the Health status of the food item is only their Macro-Nutrient quantities. Rather than just filling the missing values with Mean or Forward/ Backward Fill, I chose to run Random forest imputation, to predict the data values that are missing regarding the other Nutritional values and drawing a correlation amidst them.
- **Nutri-Score:** finally, post data preparation and augmentation in order to predict or classify this data as Healthy or Non-Healthy I needed to have a target variable to classify how well ML models are performing. For this, I leveraged the Nutri-Score technique used by Open-food-Facts [9], which revolves around a weighing system where Positive numbers are assigned to Macro-Nutrients like fibre and protein and Negative Numbers are assigned to Salt, energy, fats and finally a Nutri-Score is calculated using this algorithm from 'A'-'E'. I have labelled food items with 'A','B' as Healthy and 'C','D','E' as non-Healthy.

*3) **ML Model**:*

- Logistic Regression: One can use logistic regression for a simple classification of healthy non healthy classification of food based on their nutrient value content, as used by [8] and achieved an accuracy of 83%.
- Random Forest: Random Forest and decision trees can be leveraged to to handle non linear relationship between different features, providing feature importance score, as used by [8] and achieved an accuracy of 83%.
- Decision Tree: Decision Trees can be implemented for the classification task like [8] and achieved an accuracy of 82%
- SVM: Decision Trees can be implemented for the classification task like [8] and achieved an accuracy of 84%
- XGboost: This is a library utilized to obtain better results for our classification model using different data sampling leveraging bagging, bootstrapping and aggregating methods.

*4) **Steps**:*

- Data preprocessing: The textual data collected shall be preprocessed, this includes replacing null values by predicting using Random forest Imputation, Normalize and standardize Nutrient values and ensure correct recognition of text from image.
- Feature Engineering: The most important task is to convert data into useful format, leveraging feature engineering to formulate data in dataframes of nutrition content and standardize them to obtain values of nutrient quantity that are relative and uniform.
- Ensemble Voting: One can leverage Ensemble voting including soft and hard voting to acknowledge important factors for healthy/ non-healthy classification, like applied in [8] that gave an accuracy of 85%.
- Model training: One should ensure an appropriate train, test split and train the classifier on labelled healthy/ unhealthy data.
- Evaluation: An appropriate Evaluation metric shall be used to evaluate the performance of the classifier model. Accuracy, Recall, precision are used here.

V. Evaluation

The project was performed on a MacBookAir M1, 2020 with 8GB RAM, 245GB ssd. Here are the results for different

ML tasks performed in order to yield good results for this project.

A. OCR

- levenshtein distance: This is a string metric used to calculate difference between two sequences, as used by [8]. Here, a matrix of (M+1)*(N+1) is created and looped to understand the difference between the two words with length M and N. This shall be used to measure the edits made in the actual and predicted word by OCR.
  The Levenshtein distance between two strings a, b (of length a,b respectively) is given by Da,b where,

$$D_{a,b}(i,j) = \begin{cases} \max(i,j) if \min(i,j) = 0, \\ \min \begin{cases} D_{a,b}(i-1,j)+1, \\ D_{a,b}(i,j-1)+1, \\ D_{a,b}(i-1,j-1)+1\!\!1_{a_i \neq b_j} \end{cases} otherw \end{cases}$$

- Character Error Rate: This could calculate or keep track of the characters that are incorrectly recognized by the OCR and have to be optimized. This is the metric that helped me analyse how OCR- Text Recognition models like Tesseract, Easy-OCR and PaddleOCR did not perform well on my data.

B. Classification Model

- **Accuracy:** The proportion of correctly classified instances, Healthy or Non Healthy among all instances. The highest accuracy of 97% is achieved by Gradient Boosting Classifier.
- **Precision and Recall:** While Precision measures how a machine learning model correctly predicts the Positive Class. Recall measures how often a machine learning model correctly identifies positive instances (true positives) from all the actual positive samples in the dataset. A precision of 0.96 is yielded by Gradient Boosting Classifier and a Recall of 1.00 is achieved by Logistic Regression, SVM and Random forest.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2}$$

- **F1-Score:** This is just the harmonic mean of precision and recall, that could more effectively express the correct classification of Healthy food items. It measures how many times a model made a correct prediction across the entire dataset and is widely known in statistical analysis of binary classification and information retrieval systems.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

- **Confusion metrics:** A detailed breakdown of True positives, True negatives, False positives and False negatives is important to evaluate the correct results of the model, to check how accurate are the classifications made by the model.

Model Name	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.948276	0.919643	1.000000	0.958140
Decision Tree	0.942529	0.951456	0.951456	0.951456
Random Forest	0.959770	0.936364	1.000000	0.967136
Support Vector Machine	0.931034	0.895652	1.000000	0.944954
Gradient Boosting	0.971264	0.962264	0.990291	0.976077

TABLE I
PERFORMANCE METRICS FOR VARIOUS MODELS

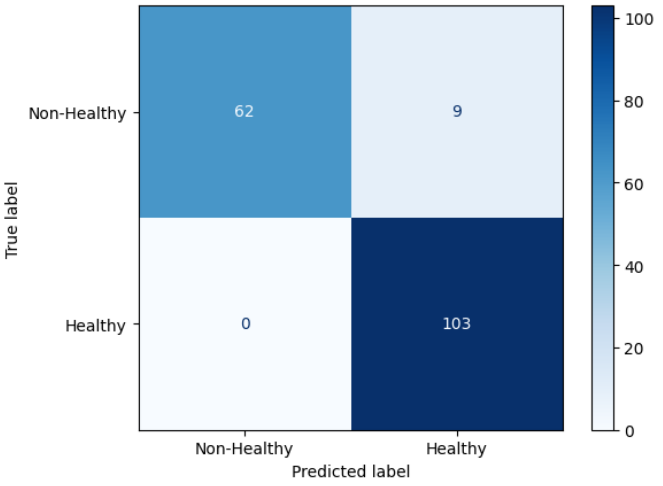


Fig. 7. Confusion Matrix: Logistic Regression Classifier

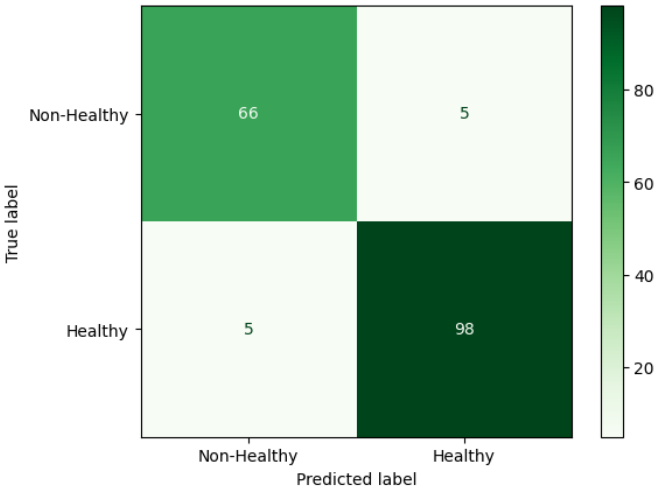


Fig. 8. Confusion Matrix: Decision Trees Classifier

- **ROC-AUC (Receiver Operating Characteristic - Area Under Curve):** A plot of the true positive rate (recall) against the false positive rate. The area under the ROC curve, representing the model's ability to distinguish between classes can be utilised to scrutinize the relationship between the two. A higher AUC indicates better performance.

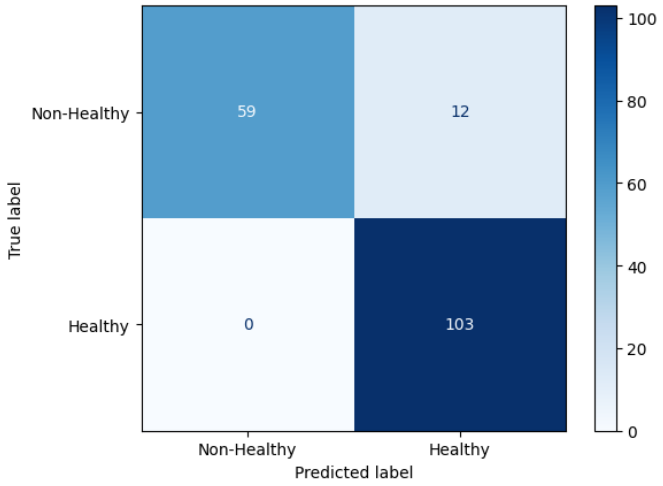


Fig. 9. Confusion Matrix: SVM Classifier

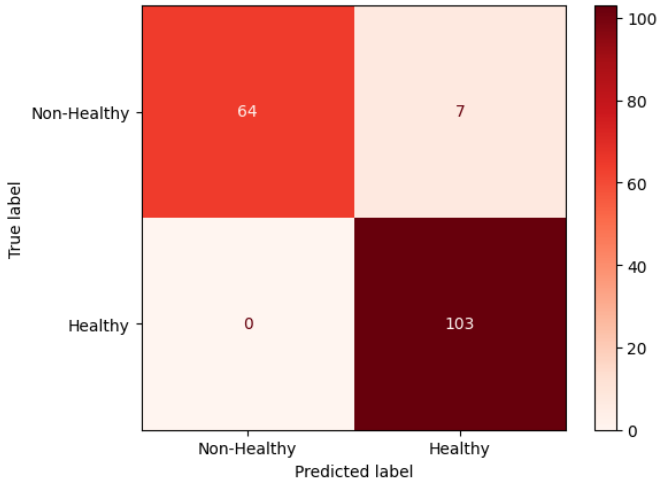


Fig. 10. Confusion Matrix: Random Forest Classifier

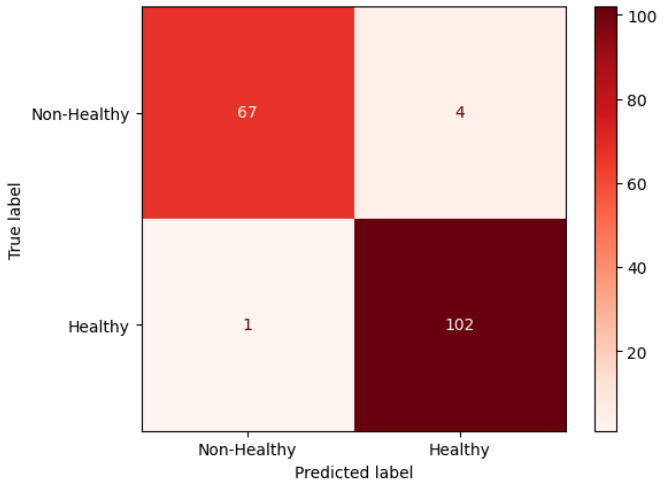


Fig. 11. Confusion Matrix: Gradient Boosting Classifier

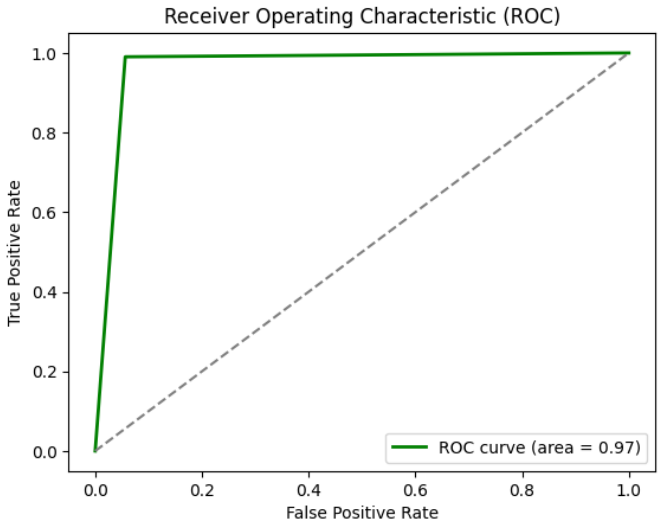


Fig. 12. ROC Curve

VI. Conclusion

In the era of abundant data amidst us we don not recognize how valuable and significant it can be and how efficiently can it be exploited for our benefit. Machine Learning plays a very vital role here, as discussed the Nutrient Content associated with packed food items is present on the packaging but we don't analyse it due to lack of time, knowledge and expertise to interpret this data efficiently. Therefore, to cater to this problem, this research has proposed a ML model that classifies a Food Product as Healthy or Non-Healthy based on its Macro-Nutrient content leveraging Nutri-Score algorithm utilised by open-food-facts [9] on a manually collected dataset of 580 Nutrient-Table images of packed food from which 347 were Healthy and 233 as Not Healthy, and were predicted by the best classifier amongst the tested ones i.e. 'Gradient Boosting' with 97% accuracy. Leveraging Google Vision API for OCR- Text Recognition and Extraction and drawing a comparison between different ML models: Random Forest Classifier, Logistic regression Classifier, decision Trees, SVM and Gradient Boosting to evaluate the best results achieved by gradient Boosting Classifier with accuracy of 97%.

VII. Future work and Limitations

This research drew its motivation from the thought of helping individuals who do not have enough time and expertise to analyse every bit of information of Nutrients and ingredients associated with the food that they buy and consume. Also with a less converged area of work with this specific purpose, to analyse the Macro-nutrient Content in a food item and classify if its Healthy for you. Although this research is a very small initial step to a bigger problem solver thee are areas that could be focused on in potential research. Gathering as much

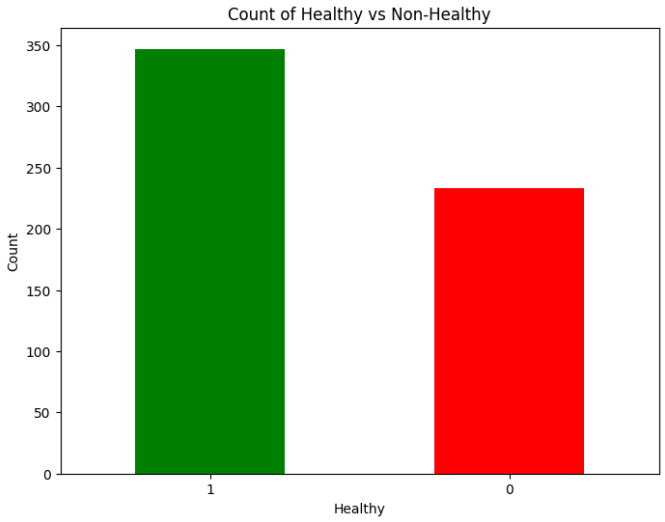


Fig. 13. Count of Healthy/Non-Healthy Food Products

data possible, limitation of dataset due to less manual power is a potential reason why I couldn't perform Neural Networks and stick to ML models for classification. Labelling is a tedious task and can be improved, since Nutri-Score is better performed when data is accurately labelled and segregated a per different genres and States or Categories. This research only takes into account the Macro-Nutrients, future work can focus on other information like Additives, allergens and ingredients to further give strong base to the predictions. Finally, there can systems developed to recommend what to eat and have a balanced diet on the history and hierarchy of an individual.

REFERENCES

[1] Mavra Ahmed et al. "Development of the Food Label Information Program: A Comprehensive Canadian Branded Food Composition Database". In: *Frontiers in Nutrition* 8 (Feb. 2022). ISSN: 2296861X. DOI: 10.3389/fnut.2021.825050.

[2] *Diabetes in Ireland*. https://www.diabetes.ie/about-us/diabetes-in-ireland/. Accessed: 2024-07-04.

[3] *EasyOCR*. https://github.com/JaidedAI/EasyOCR. Accessed: 2024-07-04.

[4] *Eating Disorders: Causes, Characteristics, and Common Misconceptions*. https://www.stpatricks.ie/media-centre/blogs-articles/2018/february/eating-disorders-causes-characteristics-and-common-misconceptions. Accessed: 2024-07-04.

[5] *Google Cloud Vision API*. https://cloud.google.com/vision?hl=en. Accessed: 2024-07-04.

[6] *HSE Obesity*. https://www.hse.ie/eng/about/who/cspd/ncps/obesity/. Accessed: 2024-07-04.

[7] S. Kayalvizhi et al. "Product Constituents and FDA regulations for Consumer Safety Using OCR". In: Institute of Electrical and Electronics Engineers Inc., 2023. ISBN: 9781665492942. DOI: 10.1109/OTCON56053.2023.10113917.

[8] Louisa Nolan. *Optical Character Recognition and Machine Learning Classification of Shopping Receipts*. URL: https://www.researchgate.net/publication/360450365.

[9] *Open Food Facts*. Accessed: 2024-08-18. 2024. URL: https://world.openfoodfacts.org/.

[10] *OpenCV*. https://github.com/opencv/opencv. Accessed: 2024-07-04.

[11] *PaddleOCR*. https://github.com/PaddlePaddle/PaddleOCR. Accessed: 2024-07-04.

[12] A. Parkavi et al. "Android application for food label recognition to ensure safe food consumption based on user allergen information leveraging OCR". In: Institute of Electrical and Electronics Engineers Inc., 2023. ISBN: 9798350335095. DOI: 10.1109/ICCCNT56998.2023.10307054.

[13] *PyImageSearch Books and Courses*. https://pyimagesearch.com/books-and-courses/. Accessed: 2024-07-04.

[14] Julia Reibring. *Photo OCR for Nutrition Labels Combining Machine Learning and General Image Processing for Text Detection of American Nutrition Labels*.

[15] Yaksh Shah. "Delving Deep into NutriScan: Automated Nutrition Table Extraction and Ingredient Recognition". In: *International Journal for Research in Applied Science and Engineering Technology* 11 (11 Nov. 2023), pp. 1596–1601. DOI: 10.22214/ijraset.2023.56852.

[16] P. Sharma and N. Verma. "The Role of Blockchain Technology in the Future of the Energy Sector". In: *Asian Finance Banking Review* 6.3 (2021), pp. 23–31. URL: https://www.afjbs.com/uploads/paper/ca485407eeb68d3a28de0916860a7f9e.pdf.

[17] *Tesseract OCR*. https://github.com/tesseract-ocr/tesseract. Accessed: 2024-07-04.