Introduction:

For people who are afflicted, their families or caregivers, the food business, and regulators, food allergies are becoming a bigger concern. Food allergy deception poses a serious threat to the food industry's image and poses a significant danger of mortality along the food supply chain. Several strategies are being investigated to address the issues, included improved food labelling and the idea of allergy symptomatic data extraction thresholds as risk management tools. These initiatives heavily rely on the capacity to accurately identify and measure food allergens, but all existing analytical methodologies have serious flaws that put the accuracy of the findings at risk, especially when it comes to the dangers of producing false positive and false negative findings.

The analytic society will have failed a major social task if we are unable to quantify food allergens recombinant and with provenance to a global unit of measurement, preventing us from realising the promise of current risk evaluation and risk mitigation of food allergens. A concerted global programme for the production of appropriately characterized clinically relevant reference books and calibrants for food allergies assessment is urgently required, along with (a) an international programme to expand the extent of proteomics and genomics bioinformatics for the genera usually contains the significant allergens to confront issues with ELISA, MS, and DNA metho-analysis. (c) the start of a global initiative that is integrated and resulting in references techniques for allergen proteins with SI-traceable findings.

In the industrialised world, food allergies—adverse immunologic (IgE and non-IgE mediated) responses to food—have caused significant morbidity5, approached epidemic levels6, and afflict up to 10% of young children and 2-3% of adults. An allergic response with a quick beginning that affects many organ systems and releases chemical messengers from basophils and mast cells, anaphylaxis, can be lethal. Even while these fatalities are relatively uncommon, their impact on allergic consumers and their families' quality of life has been researched extensively. 9–11 There are costs associated with health care, companies (product shortages, for instance), regulators, and less affluent nations where there may be serious difficulties as a result of inadequate labeling and public understanding. The only known treatments for food allergies are lifetime abstinence of the offending foods and investigational treatments (s) is required. Substantial costs are also imposed by dietary sensitivity, such as celiac disease, and stringent food restriction is frequently required.

### Legislation, risk and thresholds

Legal risk management plans for customers with allergies have centred on disclosing the existence of food allergens on labels. 15 Regulation 1169/2011, a European labeling regulation, expanded such disclosure obligations (see Table 1) to include non-prepackaged food, including those sold in catering facilities, in December 2014. 16 The basic principles of European17 and UK food law (Food Safety Act 1990) end up making it illegal to sell food that is dangerous for, or not of the essence, material, or quality imposed by allergic purchasers, especially if it was destined for their usage, may be triggered by cross-contamination with food allergies.

The hazards associated with the unintentional inclusion of allergens in food have led to an increase in cautious labelling (such as "may contain..."), which is generally viewed as inadequate. 18,19 The creation of "thresholds,"20 "action levels,"21 or "reference dosages," such as those endorsed by the European Academy of Allergy and Clinical Immunology, EAACI22 for key allergens, is a prominent area of research. Many people believe that developing evidence-based allergen mitigation techniques that are comprehended by doctors, patients, and industry is hindered by the absence of established thresholds beyond which the most sensitive allergic that individuals could respond.

For the aim of labelling, the European Food Safety Authority (EFSA) has thoroughly examined allergenic foods and food additives. However, EFSA rejected to consider supporting thresholds on the basis that labelling and the potential degree of risk (such as the percentage of the allergic community safeguarded and to what degree) are risk management considerations beyond EFSA's purview. 23 However, the main legislative provision16 controlling food labelling in the EU grants the European Commission the authority (Article 36) to enact legislation for disclosure of information on the unintended existence in food of chemicals eliciting allergy or intolerance-based responses. The allergy elicitation thresholds, which are impossible to calculate without allergen reference sources, would be the most likely foundation for such a rule. The statistics shown in the second column of Table 2 are reference dosages stated by the European Academy of Allergy and Clinical Immunology, EAACI, with the exception of fish. These data are identical to those shown in the Allergen Bureau Voluntary Incidental Trace Allergen Labeling, VITAL® scheme21.

|  |  |
| --- | --- |
| Entry | Example |
| Cereals containing gluten | Oats, Barley |
| Crustaceans | Crab. Crayfish |
| Eggs |  |
| Fish |  |
| Peanuts |  |
| Soybeans |  |
| Milk | Skimmed milk |
| Nuts | Almonds, Pecan |
| Mustard |  |
| Sesame seed |  |
| Sulfur dioxide/sulfites |  |
| Lupin |  |
| Molluscs | Oyster |
| Celery |  |

Analysis of allergens

There are several reasons why food allergen analysis is necessary. Key industry standards37 place a priority on increased accountability, tracking, and integrity in the supply chain and call for analysis to verify that food is what it is purported to be, as well as methods to limit fraudulent exposures. Analysis backs up factory cleanliness validation and verification as well as recalling and incident inquiry. 38 In order to assist and safeguard purchasers and responsible companies and to facilitate substantiation for civil or criminal action in the court system in the event of adulteration, monitoring and prosecution, notably after the advent of more substantial labeling requirements,16 heavily rely on analysis. This is a key disincentive. Analyzing unpleasant responses may be necessary in order to determine the root of the response and help the person prevent it from happening again. Examination of forensic evidence, such as food taken at the scene of the occurrence, stomach acid, or other forensic artifacts, is necessary in the already challenging26 inquiry of deaths. The cancellation of over two dozen allergy recalls on two sides of the Atlantic due to originally incorrect research raises future questions about the validity of allergen screening. This puts both the current and future evolution of allergen tolerances at danger. 39 The challenges of allergen analysis are where these issues began and where they will ultimately be solved.

**Nutrition Information**

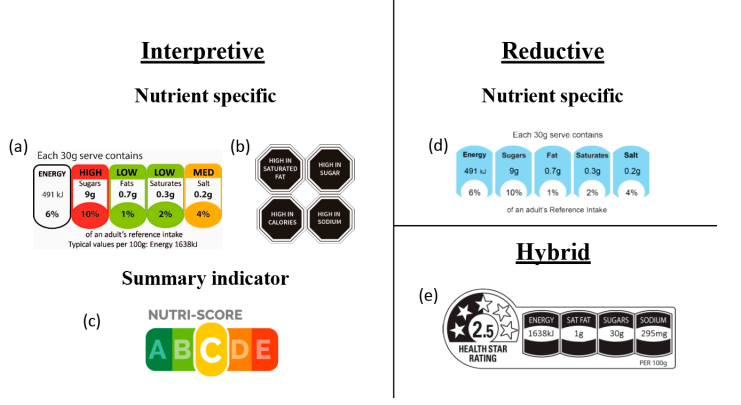
Front-of-pack labels (FoPLs) are being used as a reaction to the growing incidence of obesity in the globe [1]. pre-packaged meals are progressively being used to advise customers about the nutritional value of these items and assist them in making better decisions [2]. A substantial corpus of research backs up the idea that providing no nourishment is less successful in attaining these goals than using FoPLs On the back or side of packets, you may often find information or merely a nutrition facts panel [3][4][5]

Most of the many FoPL formats that are in use today may be categorised as reductive. either interpretative [2][6] or. Reductive FoPLs offer factual data about a food (such the levels of important nutrients inside a food) with little interpretation (like the food's role in an activity). suggested daily consumption for adults). Equivalent content, such as the levels of important nutrients, may be included in interpretive FoPLs, which similarly employ visual cues like colour to convey the food's nutritional value. Multiple traffic lights (MTL) and intakes (RI) are well-known studies of reductive and interpretative labels, respectively [7][8] Further classifications of interpretive FoPLs include nutrient-specific and summary indicator forms. While interpreting summarizing indication forms offer a general assessment of the product's nutritional quality, interpretive nutrient-specific formats (like the MTL) convey information on the particular nutrients included in a food. One further interpretative nutrient-specific type that has lately been required in a number of nations is the warning label [2]. Such FoPL

When a specified limit is surpassed, a black hexagon with the phrase "High in" followed by saturated fat, salt, sugar, or calories will frequently display. One instance of an interpretative score is the Nutri-Score. Foods are rated from A to E using a color-coded summary indication. The health star rating (HSR) similarity, which assigns meals a color-coded grade from A to E, also includes a summary indicator. In addition, the health star rating (HSR)

contains a summary indication and information pertaining to each nutrient is displayed beside the indicator,

It becomes a hybrid FoPL. Figure shows examples of these FoPLs in visual form.



Literature Review:

Feeding the world sustainably is one of our society’s grand challenges.[1](https://www.nature.com/articles/s41538-018-0021-9#ref-CR1) An exponential rise in population between 1961–2000 increased the demand for food. The demand was met by a combination of scientific and technological advances, government policy, institutional intervention and business investment, innovation and delivery. However increased farm inputs and outputs were partly at the expense of detrimental effects on the environment.[2](https://www.nature.com/articles/s41538-018-0021-9#ref-CR2),[3](https://www.nature.com/articles/s41538-018-0021-9#ref-CR3) In 2050, it is estimated there will be 9.7 billion people, and we will require about 70% more food available for human consumption than is consumed today

Allergy is one of the most important chronic diseases worldwide. It is also one of the main causes of asthma and asthma exacerbations, which has been an increasing health issue in developed countries ([Devereux, 2006](javascript:;)). Allergic hypersensitivity (IgE-type response) in sensitized individuals is elicited by allergens. The allergen–IgE interaction often results in mast cells and/or basophils releasing multiple inflammatory mediators such as histamine, leukotrienes, cytokines and chemokines. These mediators can cause a variety of symptoms from mild to severe including sneezing, itching, rashes, hives, difficulty in breathing and asthma attacks that can lead to death ([Masoli et al., 2004](javascript:;); [Stagg et al., 2013](javascript:;)).

The FAO/WHO guideline to assess allergenicity of genetically modified crops uses relaxed sequence similarity criteria. A protein is identified as a potential allergen if it harbors >35% identity with a known allergen over a window of 80 amino acids or has six contiguous amino acids that are also found in a known allergen ([FAO/WHO, 2001](javascript:;); [Metcalfe, 2005](javascript:;)). These criteria are implemented in most of the allergen databases and tools ([Mari et al., 2009](javascript:;)). However, the FAO/WHO guideline focuses on sensitivity to prevent potential new allergens entering the food market rather than accurate prediction. Therefore, these criteria yield high false-positive (FP) rates such that their application is limited ([Ladics et al., 2011](javascript:;); [Stadler and Stadler, 2003](javascript:;)). The current Codex guideline ([Codex Alimentarius Commission, 2009](javascript:;)) does not recommend the use of the six contiguous amino acid match criterion.

There are several mechanisms by which people develop adverse reactions to foods also termed [*food intolerance*](https://www.sciencedirect.com/topics/medicine-and-dentistry/nutritional-intolerance).[1](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib1) These reactions can be considered toxic or nontoxic ([Figure 1](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "fig1)).[2](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib2) Among the nontoxic reactions, those that are not immune-mediated, such as those involving [enzyme defects](https://www.sciencedirect.com/topics/medicine-and-dentistry/enzyme-defect) (eg, vasoactive amines) or reactions to certain substances (eg lactose intolerance), are far more common than immune-mediated reactions.[2](https://www.sciencedirect.com/science/article/pii/S0016508515001973#bib2) Nevertheless, immune-mediated reactions affect millions of people, are responsible for significant morbidity and health care costs, and can cause severe life-threatening reactions that lead to death.[3](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib3), [4](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib4), [5](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib5) Food allergy was defined by an expert panel of the National Institute of Allergy and Infectious Diseases as “an adverse health effect arising from a specific immune response that occurs reproducibly on exposure to a given food.” This response comprises basically all types of immune-mediated reactions, including those caused by the adaptive and innate immune system ([Figure 1](https://www.sciencedirect.com/science/article/pii/S0016508515001973#fig1)).[6](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib6)

The term allergy was coined in 1906 by the Austrian pediatrician Clemens von Pirquet,[7](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib7) who described cases of [serum sickness](https://www.sciencedirect.com/topics/medicine-and-dentistry/serum-sickness) in children treated with antibody preparations. According to Coombs and Gell,[8](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib8) there are 4 major types of allergic reactions based on pathogenesis mechanisms. The most common forms of immune-mediated adverse reactions to foods (type I reactions) always are characterized by the development of IgE against food allergens. It can be accompanied by inflammation, induced by cellular components, and mediated by T cells and eosinophils. Patients with IgE-associated food allergy can be identified based on the detection of food allergen–specific IgE in serum and body fluids, and by measuring IgE-mediated cellular and in vivo responses.[4](https://www.sciencedirect.com/science/article/pii/S0016508515001973#bib4)

Although it is tempting to speculate that food antigen–specific IgG can cause adverse reactions via type II or [type III hypersensitivity](https://www.sciencedirect.com/topics/medicine-and-dentistry/type-iii-hypersensitivity), there is no solid experimental evidence to support the relevance of these reactions to food allergies that develop in patients ([Figure 1](https://www.sciencedirect.com/science/article/pii/S0016508515001973#fig1)). Accordingly, several position papers strongly recommend against testing for food antigen–specific IgG in the diagnosis of food allergy.[9](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib9), [10](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib10)

Determinants of allergic sensitization include features of the epithelial barrier, the allergen itself (whether allergens are stable and not degraded in the environment or gastrointestinal tract), nonallergenic components of the food matrix, and substances that act as adjuvants ([Figure 2](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "fig2)A).[35](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib35) For example, food allergens have been proposed to have greater stability during digestion than other molecules in food.[36](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib36) Intrinsic factors (eg, [genetic factors](https://www.sciencedirect.com/topics/medicine-and-dentistry/heredity) such as mutations in the [*filaggrin*](https://www.sciencedirect.com/topics/medicine-and-dentistry/filaggrin) gene) and exogenous factors (eg, alcohol, anti-inflammatory drugs, pathogens, or stress) have been proposed to reduce the barrier function of the [intestinal epithelium](https://www.sciencedirect.com/topics/medicine-and-dentistry/intestinal-epithelium) and facilitate sensitization.[37](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib37), [38](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib38), [39](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib39) On the other hand, secretory antibodies, particularly [secretory IgA](https://www.sciencedirect.com/topics/medicine-and-dentistry/secretory-immunoglobulin) (SIgA), have important roles in reinforcing the epithelial barrier. Mice deficient in SIgA and secretory IgM are prone to develop food allergen–induced anaphylactic shock, which can be overcome by induction of tolerance with T-regulatory cells.[40](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib40), [41](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib41)

Many environmental and genetic factors contribute to the atopic predisposition of individuals. These determine their susceptibility to develop allergic immune responses against allergens.[42](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib42) In atopic individuals who have a predisposition toward developing IgE-associated allergies, encounters with allergen activate, after processing by antigen-presenting cells (eg, dendritic cells or B cells), allergen-specific T-helper 2 (Th2) cells, which produce cytokines such as interleukin (IL)4 and IL13. These cytokines induce class switching and production of allergen-specific IgE.[43](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib43), [44](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib44), [45](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib45) Primary allergic sensitization (such as a class switch toward IgE production) occurs early in life and leads to T-cell and IgE memory, which can be boosted with repeated allergen contact (secondary immune response).[46](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib46), [47](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib47), [48](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib48), [49](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib49) Upon contact with a primary food allergen, nonallergic individuals produce allergen-specific IgG and IgA, which do not induce allergic reactions. The formation of food allergen–specific IgE is a main feature of IgE-associated [food allergy](https://www.sciencedirect.com/topics/medicine-and-dentistry/food-allergy) and its diagnosis.

Analyses of the time courses of allergic sensitization to respiratory and food allergen sources in large birth [cohort studies](https://www.sciencedirect.com/topics/medicine-and-dentistry/cohort-analysis) have shown that food allergies and their associated symptoms develop before [respiratory allergies](https://www.sciencedirect.com/topics/medicine-and-dentistry/respiratory-tract-allergy).[50](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib50) In later life, there is a reverse trend—food allergies often are outgrown and respiratory allergies increase and dominate.[50](https://www.sciencedirect.com/science/article/pii/S0016508515001973#bib50) Interestingly, the prevalence of food allergies is approximately 10-fold lower than that of respiratory allergies.[4](https://www.sciencedirect.com/science/article/pii/S0016508515001973#bib4), [51](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib51) This could be because oral exposure to allergens activates tolerance mechanisms (via regulatory T cells) and less frequently results in allergic sensitization than respiratory exposure to allergens.[52](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib52), [53](https://www.sciencedirect.com/science/article/pii/S0016508515001973" \l "bib53)

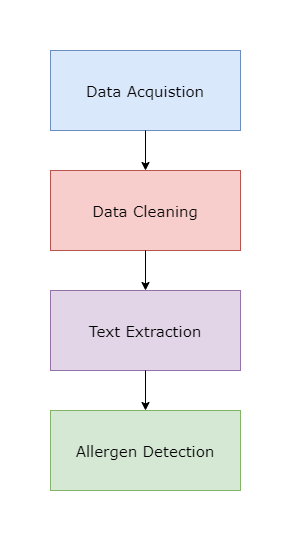
.2. Deep learning in food safety and quality evaluation Pan (2016) utilized a technique based on the hyperspectral imaging method to detect cold injury of peach, and utilized ANN to predict quality parameters. Poonnoy (2014) adopted ANN to classify the shapes of boiled shrimps based on the Relative Internal Distance (RID) values. The four shapes considered included ‘regular’, ‘no tails’, ‘one tail’, and ‘broken body’, and the RIDs were calculated by segmenting the shrimp images and drawing the co-related lines on the segmented contour. The overall prediction accuracy of the ANN classifier was 99.80%. Shafiee (2014) used MVS to capture and perform the colour transformation of the images of honey, while they used the ANN to predict key quality indices of honey (ash content, AC, antioxidant activity, AA, and total phenolic content, TPC). Zareiforoush (2015) proposed a computer vision and metaheuristic techniques to classify milled rice grains. Their method extracted the shape and size features, texture features, and colour features to build a primary feature vector, and selected the most important features based on the “greedy hill-climbing” and backtracking algorithm. The final feature vector was then used to train the ANN, SVM, Decision Trees (DT), and BN independently. The results indicated that the deep learning architecture – ANN was the best classifier. Wan (2018) designed an MVS to acquire tomato images, and the ROIs were segmented from the whole images. Finally, a Back-Propagation Neural Network (BPNN) was used to classify the maturity level of Roma tomato and pear tomato. Huang (2016) used computer vision and IR spectroscopy methods to obtain information about the organoleptic and structural changes of fish based on fish images. The PCA was employed to extract the most critical features from the data set, and a BackPropagation Artificial Neural Network (BP-ANN) was built to predict the fish freshness by training the algorithm to assess the extracted features. Khulal (2016) compared ACO (Ant Colony Optimization)-BPANN to PCA-BPANN when extracting six textural features and predict chicken quality by quantifying the total volatile essential nitrogen content in chicken. The results showed that the performance of ACO-BPANN achieved RMSEp ¼ 6.3834 mg/100g and R ¼ 0.7542, which surpassed the performance of PCA-BPANN. Zhu (2020) developed a two-layer system to grade bananas and detect the defective spots on banana peels, with an overall accuracy of 96.4%. The first layer in their system is a SVM classifier and the second layer is a YOLOv3 (You Only Look Once Version 3) model. The YOLO model is a real-time object detection model, which is capable of detecting and locating small object locations with a high speed. Pu (2015) obtained pork images via a hyperspectral imaging system in the 400–1000 nm VIS-NIR range and calibrated the images with dark and white reference images. Then, the pork muscle was segmented from the background. A variable selection method called Uninformative Variable Elimination and Successive Projections Algorithm (UVE-SPA) was adopted to find the major wavelengths with co-relations with raw spectral patterns pork categories. Histogram Statistics (HS), Gray Level Co-occurrence Matrix (GLCM), and Gray Level-Gradient Co-occurrence Matrix (GLGCM) were used to extract textural features separately, and Probabilistic Neural Network (PNN) was trained to predict the fresh and frozen-thawed level of the pork. PNN is a branch of radial basis network, which is a form of a feedforward network. It belongs to a supervised network classifier based on the Bayesian minimum risk criterion. Wu (2018) also developed a method to grade rice. The authors developed a Deep-Rice system by modifying the softmax loss function in a multi-view CNN and applied the CNN model to extract the discriminative features from several angles of the rice. Additionally, the authors built a data set of rice – FIST-Rice. With 30,000 rice kernel samples, this data set can be used to research food security .2.3. Deep learning in food process monitoring and packaging Teimouri (2018) extracted colour, geometrical aspects, and texture features with a CCD camera from chicken portions, then adopted the PLSR, LDA, and ANN to classify the data to achieve on-line separation and sorting of the food product. The conveyor that carried the food moved at 0.2 m/s, and it took 15 ms to process one image. The system’s overall accuracy was 93%, and 2800 food samples could be sorted in 1 h. De Sousa Ribeiro (2018a) offered an adaptable CNN-based system to identify the missing information on food packaging labels. In this research, K-means clustering and KNN classification algorithms were used to extract the data’s centroids and provided them to CNN. Also, De Sousa Ribeiro (2018b) proposed an end-to-end deep neural network for detecting the dates on food packages. Optical Character Verification (OCV) and Recognition (OCR) (which are methods to convert printed characters into computer text) were used to extract the ROI and feed the data to a Fully Convolutional Network (FCN) to locate the data. Then a Maximally Stable Extremal Regions (MSER) was adopted to segment the data. FCN solves semantic segmentation by pixel-level classification of images, and MSER is used for spot detection in an image. 4.2.4. Deep learning in foreign object detecting Pujari (2013) proposed a system to detect and classify fungi contamination on commercial crops. The authors applied Discrete Wavelet Transform (DWT) and PCA to extract significant features from crops images and feed them to Mahalanobis distance and PNN classifiers. The final prediction accuracy of these two classifiers was 83.17% and 86.48%, respectively. Ma (2017) carried out a method to segment greenhouse vegetables with foliar disease spots in the real field. The authors adopted Comprehensive Color Feature (CCF), including colour information Excess Red Index (ExR), H component of HSV colour space and b\* component of L\*a\*b\* colour space, as well as region growing, to segment disease spots from clutter backgrounds. This method overcame the issues of uneven illumination and clutter background encountered in the real field. The CCF were then sent to a CNN architecture to classify the spots. The precision of this algorithm achieved an accuracy of 97.29%. Dos Santos Ferreira (2017) developed software to segment and detect weeds growing among soybean crops. This software classification function is based on a CNN trained with crop images captured in real fields by three professional drones. The system adopted the superpixel segmentation algorithm - Simple Linear Iterative Clustering (SLIC), which groups the pixels into atomic regions to replace the pixel grid, and to segment the undesirable weeds from soil and soybean crops. Ebrahimi (2014) proposed a system to distinguish wheat grains from weed seeds, with the goal of determining the purity and grade of the wheat seeds. The authors used a hybrid of Imperialist Competitive Algorithm (ICA) (which is used for optimization problems) and ANN as classifier, to identify the optimal characteristic parameters. Al-Sarayreh (2019) proposed a foreign object detection system for meat products. This system used real-time hyperspectral imaging to extract both spectral and spatial features of the target food, integrating it with a sequential deep-learning framework that is composed of region-proposal networks (RPNs) and 3D-CNNs. Rong (2019) developed two CNNs. These two models can detect foreign objects such as flesh leaf debris and metal parts in walnuts and segment the foreign objects

There were two components of the project

1. Allergen detection
2. Nutrition score and nutrition band prediction

**Allergen Detection**

The high level overview of the diagram is shown here,



Data Acquisition:

Data Cleaning:

Text Extraction:

As it was an image dataset, the text from the image was extracted to detect the allergens present in the ingredients. In deep learning, the problem of text extraction is called ‘Optical Character Recognition (OCR). It has two main components. The first component is detecting the text and the second is transcribing it.

**Text detection:**

It is the task of locating the text in the image or page. The image is represented as a three dimensional vector, depicting height, width and number of channels i.e. H,W,C. the channels represent RGB colors. Detecting the text is a difficult task due to number of reason. The text in the image comes in various shapes. Moreover, the orientation can be distorted. In the case of image, the problems increase because the quality of image can be poor, the background color can be too dark to detect text and the way the image is captured. The researchers pose the text detection problem as object detection or instance segmentation. If the text detection problem is posed as the object detection problem, the model learns to detect output coordinates of bounding boxes around text. On the other hand, if the text detection problem is posed as the instance segmentation task the model learns to produce mask in which the pixels having text is marked and pixels without text are not marked.

**Detecting text as an Object Detection problem**

Detecting the text has mainly relied on manually creating features to identify characters [9][10]. Deep learning improvements, particularly in object identification and semantic text detection has been approached differently as a result of segmentation. Researchers created effective text detection techniques by utilizing these high-performing object detectors from the classic computer vision research, for example the Single-Shot MultiBox Detector (SSD) and Faster R-CNN models [11][12]

TextBoxes is one of the earliest works to use a regression-based text detection method [13]. In order to adjust the object detection model to text, they introduced lengthy baseline boxes with big dimensions to SSD. This study was the foundation for many articles that developed the Deep Matching Prior Network (DMPNet) and the Rotation-Sensitive Regression Detector (RRD) to create regression-based algorithms robust to orientations [14][15]. Similar methods are used in other articles, but they create their unique proposal networks that are tailored around texts instead of real images. For example, [16] use a vertical anchor mechanism in their Connectionist Text Proposal Network to integrate convolutional networks with recurrent networks to enhance accuracy for detecting text which is horizontal.

In most cases, the intersection over union (IoU) metric and an F1 score are used to assess object identification models. The metric determines how much of a prospective bounding box intersects with the original bounding box (the intersection) and divides that amount by the entire amount of space that both the candidates and ground truth bounding boxes take up (the union). The following step is to select an IoU limit to identify which anticipated boxes are true positives (IoU ). False positives are what's left of them. A false negative is a box that the algorithm is unable to identify. These parameters are used to calculate an F1 score to assess the object detection system.

**Detecting text as an instance segmentation problem:**

Text identification in written documents presents a special set of difficulties since it is frequently thick and contains a lot more text than is typically seen in natural photos. Text detection can be set up as an extremely dense instance segmentation operation to get around the density issue. The process of categorizing every pixel in a picture into a set of predetermined classes is known as instance segmentation.

Detection that is based on segmentation locate text areas at the pixel level. In a unified framework, these per-pixel predictions are frequently utilized to estimate probability of text sections, characters, and their interactions with one another. When text is deformed or mismatched, professionals employ widely used segmentation techniques like Fully Convolutional Networks (FCN) to recognize text [17], outperforming object identification approaches. By explicitly obtaining bounding regions from the segmentation result, several works expand on this segmentation basis to provide word bounding regions [18][19]. By extrapolating the text area, centre line, text direction, and potential radius from a Fully connected layers, TextSnake goes one step further [20]

**Transcribing the text**

Transcribing the text that appears in images is known as text transcription. The inputs, which are often images, have the dimensions C H W0 and corresponds to either a character, word, or string of words. These cropped images must be input into a model designed for transcribing the text, which must then learn to generate a string of tokens from some predetermined vocabulary V. V often translates as a string of characters. This is the most logical method, for example, for recognizing digits [21]. Otherwise, V can also represent a collection of words, much like an issue involving word-level language modelling [22]. In both situations, the problem may be expressed as a classification problem with multiple classes. The size of vocabulary V determines the total number of classes in the model.

Since there are more classes than characters in the multi-class classification, word-level text transcription models need additional data. Anticipating words rather than characters reduces the likelihood of committing minor errors (like replacing "g" by "9" in a word for example “bag”). However, if one is restricted to a word-level vocabulary, it is impossible to transcribe terms that are not included in this vocabulary. Since there are a finite amount of characters, this issue does not arise at the character level. Building a vocabulary that includes all the letters is simple as much as we are familiar with the article's languages.

Recurrent neural networks (RNNS), particularly recurrent models using Long short term memory (LSTM) or gated recurrent unit (GRU) on top of the feature extraction done by using convolutional neural network, have recently gained popularity in the scientific community [23] [24]. Two distinct decoding techniques are frequently applied to the transcription of a token. One technique is the utilization of beam search with cross entropy loss and attention-based sequence decoder. Howevr, the efficiency of conventional sequence attention can be decreased by visuals that are occasionally incorrectly aligned or badly positioned.

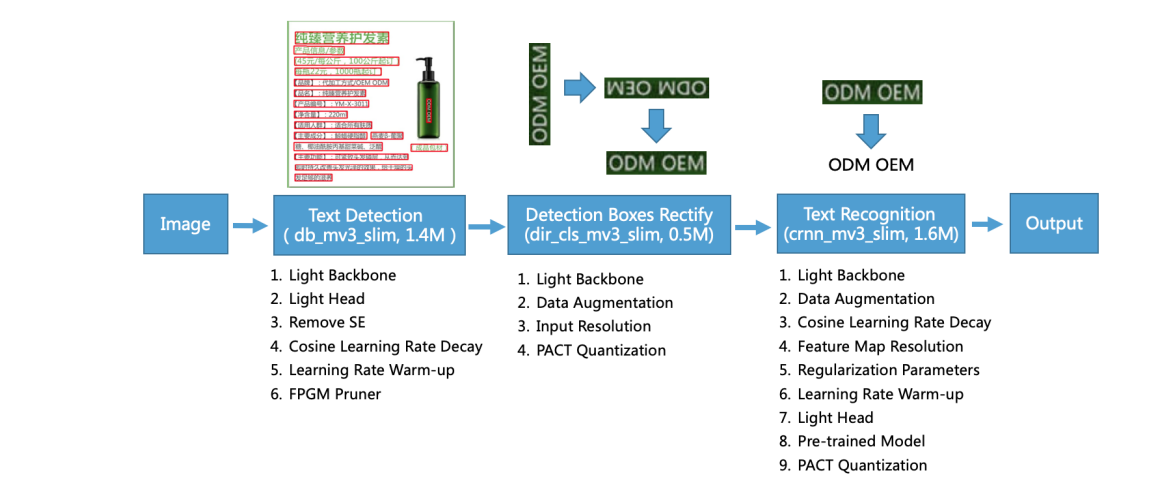
[25] employed spatial attention methods explicitly, whereas [26] used attention alignment, straightforwardly storing spatial information of letters. A typical loss function in speech that accurately represents repetitive characters in sequential outcomes is connectionist temporal classification (CTC) loss, which is used often for transcriptional decoding [27]. The bulk of textual transcribing models draw on developments in sequence modelling for both voice and text, and frequently only require small tweaks to make good use of these developments. In contrast to the other elements of the document interpretation job, professionals rarely explicitly address this issue.

Since the dataset contained food packaging images, two pretrained OCR models were used.

1. Paddle OCR
2. Keras OCR

**Paddle OCR:**

A useful, extremely lightweight OCR system is PP-OCR. Only 3.5M and 2.8M of the PP-total OCR's model are needed to recognize 6622 Chinese symbols and 63 alphanumeric symbols, respectively. A text detector (using 97K images), an orientation classifier (using 600K images), and a text recognition system are among the pre-trained models that are available for Chinese and English recognition (17.9M images are used). Additionally, the proposed PP-OCR have been tested in French, Korean, Japanese, and German language recognition tasks.



**Detecting text**

Finding the text area in the image is the goal of text detection. We employ Differentiable Binarization (DB), a text detector in PP-OCR that is based on a straightforward segmentation network [28]. Because of its straightforward postprocessing, DB is particularly effective. The following six tactics are applied to further enhance its efficacy and efficiency: light backbone, light head, remove SE module, cosine learning rate decline, learning rate warm-up, and FPGM pruner. The text detector's model size is finally down to 1.4M.

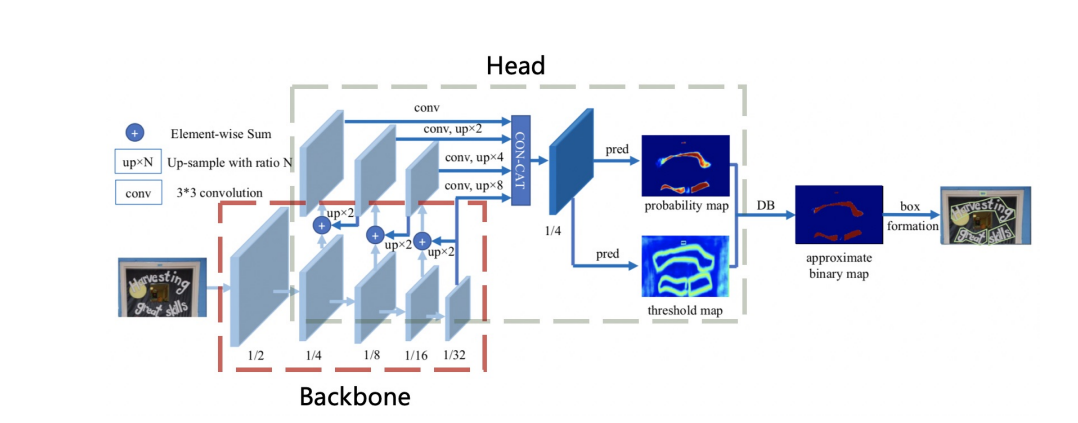
**Rectification of Detection Boxes**

Since the identification box is made up of four points, it is simple to create via geometric transformation the text box's transformation into a horizontal rectangular box for future text recognition. The repaired boxes, however, may be turned around. In order to establish the text orientation, a classifier is required. Additional switching is necessary if a box is found to be flipped. A straightforward picture classification problem is training a text direction classifier. To improve the model's functionality and shrink the model size, we use the four accompanying strategies: light backbone, data augmentation, input resolution, and PACT quantization. The text direction classifier's model capacity is 500KB in total.

**Text Recognition**

For text recognition in PP-OCR, we employ the well-known and useful CRNN (Shi, Bai, and Yao 2016) text recognizer. CRNN combines sequence modelling and feature extraction. To prevent the discrepancy among forecast and categorization, it utilises the Connectionist Temporal Classification(CTC) loss. The following nine strategies—light backbone, data augmentation, cosine learning rate decay, feature map resolution, regularisation parameters, learning rate warm-up, light head, pre-trained model, and PACT quantization—are used to improve model performance and shrink model size in text recognizers. Last but not least, the text recognizer's model size is just 1.6M for English and Chinese identification and 900KB for letters and numbers character identification.

We build a huge dataset for Chinese and English recognition as an example to design a viable OCR system. There are 97K photos in the text detection dataset specifically. 600k pictures make up the direction classification dataset. A 17.9M picture text recognition dataset is available. To execute ablation tests fast and pick the best tactics, a small sample of the data is chosen. In Figure 2, we conduct several ablation experiments to demonstrate the outcomes of various tactics. Additionally, we test the suggested PP-OCR system's ability to recognise various languages, such as those using alphanumeric symbols, French, Korean, Japanese, and German.



**Enhancement or Slimming Strategies**

**Text detection**

**Light Backbone**

The size of a text detector's model is mostly influenced by the backbone's size. Hence, while creating super lightweight frameworks, light backbones should be chosen. MobileNetV1, MobileNetV2, MobileNetV3, and ShuffleNetV2 series are frequently utilised as the light backbones in the establishment of picture categorization. The scale varies for every series. Figure 6 illustrates that certain light backbones, such as the MobileNetV1, MobileNetV2, MobileNet3 and ShuffleNetV2 series, performed on the ImageNet 1000 classification thanks to the inference time on CPU and accuracy of PaddleClas1.

On the Snapdragon 855 (SD855), the induction duration is examined with the batch size set to 1. When the predict time is the same, MobileNetV3 may produce more accuracy. In terms of size selection, we use MobileNetV3 big x0.5 to experimentally balance accuracy and efficiency. In addition, PaddleClas offers 122 models' pretrained weights and their assessment methods, as well as up to 24 series of image categorization connectivity frameworks and training configurations. These models include ResNet, ResNet vd, SEResNeXt, Res2Net, Res2Net vd, DPN, DenseNet, EfficientNet, Xception, HRNet, etc.

**Light Head**

The core of the textual detection is akin to the FPN [29] architecture in object identification and fuses the feature maps of various scales to enhance the effect for the recognition of tiny text areas. One-to-one convolution is frequently used to compress feature maps to the same number of channels for the simplicity of combining the various resolution feature maps (we utilise inner channels for this). From the fused feature map with convolutions that are also connected to the aforementioned inner channels, the probability map and threshold map are produced. Thus, interior channels have a significant impact on model size. The model size is decreased from 7M to 4.1M by reducing the inner channels from 256 to 96,

although there is a minor loss in precision.

**Remove SE**

The abbreviation SE stands for squeeze-and-excite [30]. SE blocks explicitly represent channels interconnections and flexibly re-calibrate channel-wise feature responses. Since it is evident that SE blocks can increase the accuracy of visual tasks, MobileNetV3's search space includes them, and the architectural style of the network includes a large number of SE blocks. Nevertheless, it is challenging to estimate the channel-wise feature responses using the SE block when the input resolution is high, such as 640 640. Although there is little accuracy gain, the time investment is substantial. The model size is decreased from 4.1M to 2.5M when the SE blocks are removed from the backbone, but the correctness is unaffected.

**Cosine Learning Rate Decay**

The hyperparameter that controls learning speed is learning rate. The loss value changes more slowly the slower the learning rate. Although a low learning rate might help to guarantee that you don't miss any local minimums, it also slows convergence. Because the weights are first initialised randomly, we may specify a reasonably high learning rate to facilitate rapid convergence. As training progresses, the weights become more near to their ideal values, necessitating the adoption of a comparatively slower learning rate.

**Learning Rate Warm-up**

According to the study [31], applying learning rate warm-up operations can assist to increase picture classification accuracy. It is advised to use a lower learning rate at first since employing a big learning rate too early on might cause numerical instability. The initial learning rate should be applied once the training process has stabilised. The results demonstrate that this method is also successful for text detection.

**FPGM Pruner**

Another strategy to increase the effectiveness of a neural network model's inference is pruning. We employ FPGM [32] to identify the insignificant sub-network in the original models in order to prevent the model performance decrease brought on by model reduction. Each filter in a convolution layer is regarded as a point in Euclidean space and the geometric median is used as the FPGM's criteria. Then, as illustrated in Figure 9, compute the geometric median of these points and eliminate the filters with identical values. For model pruning, each layer's compression proportion is indeed crucial. Regularly, homogeneous layer pruning results in a considerable productivity hit. In PP-OCR, the approach described in [33] is employed to compute every layer's pruning susceptibility, which is then utilised to determine how redundant every layer is.

**Direction Classification**

This section will go into depth about four techniques for improving a direction classifier's model capability or shrinking its model size.

**Light Backbone**

The orientation predictor, which uses the same MobileNetV3 as the textual detection, also uses this technology as its foundation. We experimentally balance accuracy and efficiency using MobileNetV3 tiny x0.35 since this work is reasonably straightforward. The precision does not increase more often with bigger backbones.

**Data Augmentation**

In this study [33] image processing techniques such rotation, perspective distortion, motion blur, as well as Random noise are demonstrated. These procedures are collectively known as BDA (Base Data Augmentation). They are inserted at random to the practise pictures. The experimentations illustrates that BDA is helpful for training the direction classifier as well. For example, CutOut (DeVries and Taylor 2017), RandErasing (Zhong et al. 2020), HideAndSeek (Singh and Lee 2017), GridMask (Chen 2020), Mixup (Zhang et al. 2017), and Cutmix are a few novel data augmentation operations that have been lately suggested in addition to BDA for enhancing the impact of categorization tasks (Yun et al. 2019). However, the trials reveal that, with the exception of RandAugment and RandErasing, the most of them are useless for training direction classifiers. It works best with RandAugment. Finally, we enrich the training data for the direction categorization with BDA and RandAugment.

**Input Resolution**

Generally, accuracy will grow as input size of a normalised image increases. Since the direction classifier's core is so lightweight, appropriately raising the resolution won't result in a noticeable increase in calculation time. The majority of earlier text recognition techniques set a normalised image's height and width to 32 and 100, respectively. The width and height are adjusted to 48 and 192, respectively, in PP-OCR in order to increase the orientation classifier's accuracy.

**PACT Quantization**

The neural network model can have reduced delay, density, and computing resource utilization thanks to quantization. The two primary categories of quantization at this time are offline quantization and online quantization. Standalone quantization is a fixed-point quantization technique which does not need retrain and instead determines the quantization parameters using techniques like KL divergence and moving average. Throughout the training phase, quantization parameters are determined using online quantization, that might result in less quantization loss than offline quantization mode. A novel online quantification technique called PACT (PArameterized Clipping acTivation) eliminates certain anomalies from the activations beforehand (Choi et al. 2018). The model can learn more suitable quantitative scales after the outliers have been eliminated. The following is the PACT preprocessing equation for activations.

The ReLU function provides the foundation for the preprocessing of the activation value of the standard PACT approach. Any activation rates that are higher than a specific limit are trimmed. However, MobileNetV3's activation features include both hard swish and ReLU activation. A larger quantization loss results from using standard PACT quantization. To lessen the quantization loss, we thus change the activations preprocessing formula to the following. Figure 10 shows the textual recognizer's CRNN framework. This number is taken from the publication (Shi, Bai, and Yao 2016). The text recognizer's head and backbone are depicted individually in the red and grey rectangles.

The ReLU function provides the foundation for the preprocessing of the activation value of the standard PACT approach. Any activation rates that are higher than a specific limit get trimmed. Nevertheless, MobileNetV3's activation features include both hard swish and ReLU activation. A larger quantization loss results from using standard PACT quantization. To lessen overall quantization cost, we thus change the activations preprocessing formula to the following. Figure 10 shows the textual recognizer's CRNN framework. This number is taken from the publication (Shi, Bai, and Yao 2016). The text recognizer's head and backbone are depicted individually in the red and grey rectangles.

**Text Recognition**

In this part, nine ways for improving the model capability or condensing the model size of a textual recognition system will be described in depth. The layout of the textual recognition system CRNN is seen in Figure 10.

**Light Backbone**

As with text detection, MobileNetV3 serves as the foundation of our text recognizer as well. For experimentally optimize accuracy and efficiency, MobileNetV3 tiny x0.5 is used. MobileNetV3 tiny x1.0 is also a viable option if you're not very concerned about model size. Just by adding 2M to this model, its accuracy has visibly risen.

**Data Augmentation**

TIA (Luo et al. 2020) is another successful data augmentation technique for text recognition in addition to BDA (Base Data Augmentation), which would be frequently employed in script identification as previously discussed. An initial set of fiducial points is established on the picture, as seen in Figure 11, initially. To create a new picture using the geometric transformation, change the points at random. In PP-OCR, we supplement the text recognition training pictures using BDA and TIA.

**Cosine Learning Rate Decay**

The optimal learning rate lowering technique is cosine learning rate degradation, as was discussed in text detection. The results of the trials demonstrate that the cosine learning rate decay technique may also be used to improve the model's text categorization capabilities.

**Feature Map Resolution**

In PPOCR, the height and breadth of the CRNN input are set to 32 and 320 respectively in order to accommodate multilingual identification, notably in Chinese recognition. The original MobileNetV3's advancements are so inappropriate for textual identification. Figure 12 illustrates that researchers modified the down sampling feature map's duration, but for the first one, which was changed from (2,2) to (2,1). We also change the stride of the second down sampling feature map from (2,1) to maintain additional vertical knowledge (1,1). As a result, the resolution of the entire feature map as well as the precision for the textual recognition system are dramatically impacted by the stride of the second down sampling feature map, or s2. To experimentally attain the increased quality in PP-OCR, s2 is set to (1,1).

**Regularization Parameters**

A prevalent phrase within machine learning includes overfitting. The algorithm works well on the training data, but it performs poorly on the test data, to put it simply. Numerous conventional methods have been developed to prevent overfitting. Weight decay is one of the most popular methods to prevent overfitting amongst them. L2 regularisation (also known as L2 decay) is applied to the loss function after the final loss function. The weight of the network tends to select a lower value with the aid of L2 regularisation, and ultimately all of the network's parameters converge to zero, improving the model's generalization capability. L2 degradation significantly affects textual recognition performance.

**Learning Rate Warm-up**

Warming up the learning rate helps with textual detection in a manner akin to textual identification. The results indicate that employing this technique is also successful for textual detection.

**Light Head**

The sequence characteristics are encoded to the conventional anticipated characters using a complete connecting layer. The scale of a textual recognizer's model is influenced by the dimension of the sequence characteristics, particularly for Chinese recognition where there are more than 6 thousand of symbols. However, it is not true that the capacity to express sequence characteristics is stronger the greater the dimensionality. The dimension of the sequence features in PP-OCR is experimentally fixed at 48.

**Pre-trained Model**

The current networks, that were learned on a huge data set like ImageNet, can be fine-tuned if the training data is sparse in order to obtain quick closure and improved accuracy. The effectiveness of the aforementioned technique is demonstrated through transfer learning for object identification and picture categorization. The amount of data utilised for text recognition in actual settings is frequently little. The accuracy may be considerably increased using the aforementioned approaches if the models are trained on tens of millions of examples, including synthetic ones. Through tests, we show how successful this tactic is.

**PACT Quantization**

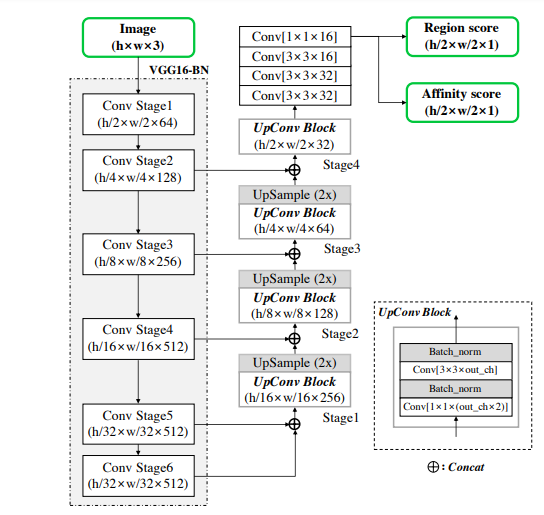
With the exception of omitting the LSTM layers, we use a quantized strategy to minimize the model size of a text recognizer that is similar to that used for direction classification. These layers won't be measured.

**Keras OCR**

Keras OCR was based on CRAFT (Character Region Awareness For Text detection”). It locates the various character areas and connects the discovered characters to a text occurrence. The character region score and affinity score are produced by a convolutional neural network in the system,. Every character in the given input image is localized using the area score and grouped into a single instance using the affinity score. We provide a minimally supervised learning approach that assesses character-level ground truths in current actual word-level datasets in order to make up for the absence of character-level labels.

**Network Architecture:**

The foundation of CRAFT is a fully convolutional network architecture based on VGG-16 [33] with batch normalization. In the decoder portion of the model, skip connections are used, which is comparable to U-net [30] in that it accumulates low-level characteristics. The affinity score and region score are the two score maps that make up the final result. Following figure depicts the system layout conceptually.



**Training the model:**

**Producing the ground truth labels**

The character level bounding box coordinates are created for the ground truth label in order to calculate the region score and affinity score for given input image. The likelihood that a particular pixel is the character's centre is represented by the region score, and the likelihood that a provided pixel is in the middle of a gap between two characters is represented by the affinity score.

Contrary to a binary segmentation map, that individually classifies every pixel, authors used a Gaussian heatmap to express the likelihood that the character centre will appear. Because of its outstanding versatility when working with regions of ground truth that are not rigorously delimited, this heatmap approach has been applied in various areas, such as pose estimation works [1, 28]. The region score and affinity score are both learned using the heatmap depiction.

It takes a long time to explicitly calculate the Gaussian distribution value for every pixel inside the bounding box. authors utilized the following techniques to roughly create the ground truth for calculating both the affinity score and region score because character bounding boxes on a picture are frequently warped by viewpoint estimates: create a 2-dimensional isotropic Gaussian map; calculate the perspective transform between every character box and the area of the Gaussian map; 3) distort the box region on the Gaussian map. The affinity boxes are formed utilizing neighbouring character boxes for the ground facts of the affinity score. Two triangles were created, that were called as the lower and upper character triangles, by plotting diagonal lines to link the edges of every character box that are opposite one another. Then, an affinity box is created for each pair of adjacent character boxes by assigning the top and bottom triangles' centres as the box's edges.

Despite employing limited receptive fields, the suggested ground truth formulation allows the model to recognise huge or long-length text occurrences appropriately. In contrast, in such situations, earlier methods like box regression call on a huge receptive field. Convolutional filters may now concentrate exclusively on inter-character and intra-character information rather than the complete text sample due to the employed character-level detection method.

**Weak Supervised learning:**

Real input images in a dataset typically include word-level labels, in contrast to synthetic datasets. Here, we minimally oversee the generation of character boxes from every word-level label. The learnt intermediate model forecasts the character region score of the clipped word pictures to produce character-level bounding boxes whenever a real image with word-level labels is shown. The score of the trust map across every word box is calculated proportionally to the count of identified characters divided by the count of ground truth characters, which is utilized for the learning weight throughout training, in order to indicate the accuracy of the intermediate model's projection. The raw images are first trimmed to create the word-level graphics. Secondly, the region score is predicted by the model as of late. Finally, the character areas are divided using the watershed technique [35] to create the character bounding boxes encompassing regions. The reverse transform from the trimming stage is then applied to convert the coordinates of the character boxes back into those of the original picture. The procedures shown in Fig. 3 may be used with the acquired quadrilateral character-level bounding boxes to construct the pseudo-ground truths (pseudoGTs) for the region score and the affinity score. Researchers trained utilizing partial pseudo-GTs when the model has been trained employing weak-supervision. The result may be fuzzier inside character areas if the model was trained using erroneous region scores. To avoid this, they evaluated every one of the model's produced pseudo-GTs. Luckily, the word length in the textual labels serves as a very powerful indication. The majority of datasets include word transcriptions, and the size of the words may be used to gauge how confident the pseudoGTs are.

Let R(w) and l(w) be the bounding box area and word length, respectively, for a word-level labeled sample w of the training examples. We can determine the predicted character bounding boxes and their related character lengths lc(w)using the character splitting method sconf (w). The reliability score for the instance w is then calculated as,

Where the image's pixel-wise confidence map Sc is calculated as,

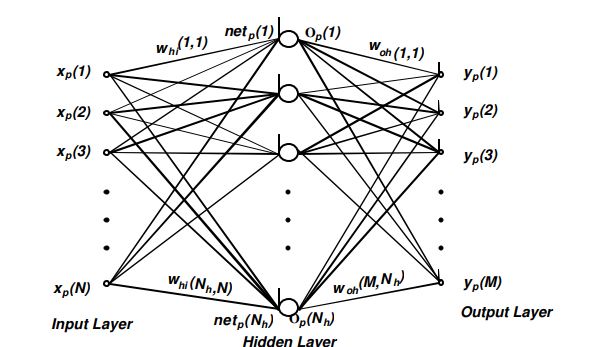
where p stands for a pixel inside area R. (w). The definition of the objective L is

wherein and stand for the projected region score and affinity score, respectively, and and stand for the pseudo-ground truth region score and affinity map, respectively. Since we can retrieve the actual ground truth while training using synthetic data, is set to 1. The CRAFT model's ability to forecast characters improves with training, and the certainty ratings sconf(w) rise along with it. The character region score map throughout training. The area scores are initially rather low for text that is unknown in natural images.

The algorithm picks up on novel text patterns like erratic typefaces and synthesised texts with distinct data distributions from the SynthText dataset. The predicted character bounding boxes should be disregarded since they have negative impacts on model training if the certainty score sconf (w) is less than 0.5. By merely dividing the word region R(w) by the number of characters l, we may calculate the character-level estimations in this scenario by assuming that each character's breadth is constant (w). Then, sconf (w) is adjusted to 0.5 to acquire texts' hidden meanings.

**Multi-layer Perceptron:**

Over time, MLPs have become a highly potent method for addressing a wide range of issues. Both the performance of artificial neural systems and our comprehension of how they function have improved significantly. Multi-layer Perceptron constitute units that are arranged in layers, every layer in MLP is contains nodes. In the case of fully connected network, every node is connected to every other node in the subsequent layers. every Multi layer perceptron model contain three layers at least. One layer is input layer, one or more hidden layers and the final layer i.e. output layer. The input to the model is passed through the input layer. This layer takes the input and passes it to the next layers i.e hidden layer. The nodes in input layer haave linear activation functions without any threshold. Every node in the hidden layer and in the output layer however contains thresholds in addition to the weights. in contrast to nodes in input layer, the hidden layer nodes have nonlinear activation functions. The nodes in output layer however contains the linear activation function. this, every signal that is fed into the node of subsequent layer has the initial input multiplied by the weight and threshold is added. This is then passed to the activation function that will be linear in case of output layer and nonlinear in hidden layers. a three layered MLP is shown in figure.



The training datset contains a set of Nv patterns (xp, tp)., here t denotes the number of pattern. xp is denoting the N-dimensional input vector of a training pattern p. in the figure, yp represents the output vector that is obtained after training the MLP model with the input xp. for providing the convience thresholds on the nodes of hidden layer and output layers are shown by giving the value of one to the augemented vector component. It is denoted by xp(N+1). As mentioned earlier the input and output layer has linear activation functions. the input to the hidden unit is expressed formally as:

The output function Op(j) for the training pattern p can be expressed as:

Conventionally the sigmoid activation function is used as a nonlinear activation function.

The performance of MLP is usually measured by mean square error. Formally,

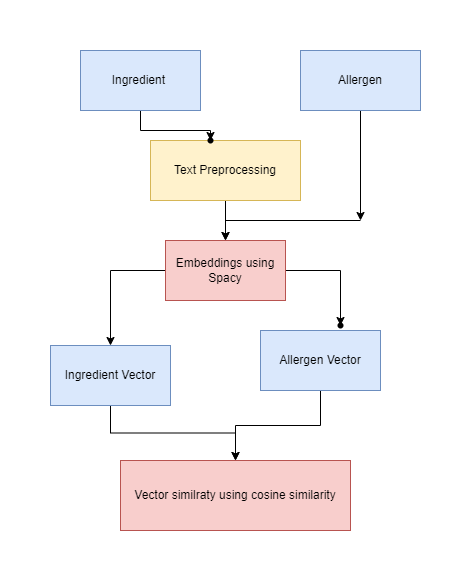
Where,

The term Ep denotes the error of the pth training pattern. Here is desired output for the pattern. In addition, it permits the calculation of mapping error for ith output. Mathematically,

The ith output is expressed as:

Here denotes the weights from the nodes of input layer to the nodes of output layer and denotes the weights from the nodes of hidden layer to the nodes of output layer

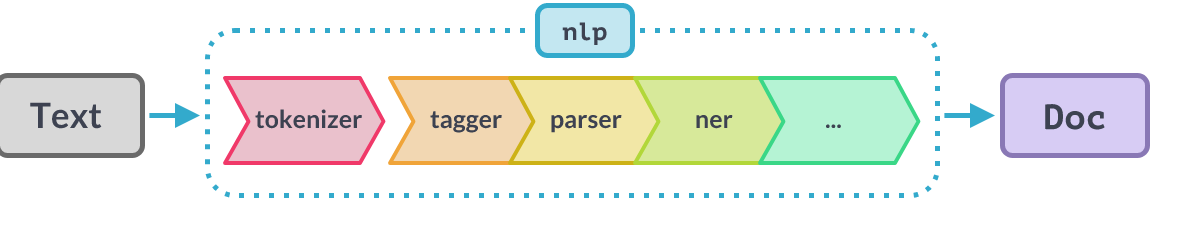
**Allergen detection pipeline**

****

**Word Similarity**

The allergen detection was treated as the word similarity problem. After extracting the text from images using OCR, the next step was to convert the text into vector form. the OCR extracted ingredients were converted into vectors and these vectors were then compared with the vectors of allergens. We selected 250 allergens and transformed them into vectors. The cosine similarity was then measured between OCR extracted ingredient and all the allergen vectors. If any ingredient and allergen get the cosine similarity more the 90, that ingredient is declared as the allergen.

In order to convert text into vectors, spacy library was used. it provides a function called nlp. This function first tokenize the text if the input is sentence, then the processing pipeline is employed. The text is tagged using tagger, lemmatized using lemmatizer, parsed and finally entity recognizer is used. the final output is a vector of the word.



There were other alternatives available to convert text into vectors for example BERT and word2Vec. However, both the models embed the words according to the context. The word embedding or text representation is produced by considering the context and semantics of word. For example, if the sentence ‘I like books’ is passed, it will produce similar embeddings for like and book. but we were not dealing with the sentences. Our problem was simply to get the vector representation of a single word that would be compared with another word to see the similarity. So, we did not need contextual or semantic word embeddings.

**Cosine similarity:**

The ingredient text and allergen text was converted into vectors. Let the ingredient text denoted with i and the allergen text denoted as a.

The and are the vectors of ingredient and allergen respectively. The cosine similarity between allergen and ingredient is given as:

Every ingredient extracted from the package image using OCR was compared with every allergen in the database. the ingredient vector that was more then 90% similar with allergen vector was printed as allergen.

[1] N. C. D. R. F. Collaboration, “Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19· 2 million participants,” *Lancet*, vol. 387, no. 10026, pp. 1377–1396, 2016.

[2] R. Kanter, L. Vanderlee, and S. Vandevijvere, “Front-of-package nutrition labelling policy: global progress and future directions,” *Public Health Nutr.*, vol. 21, no. 8, pp. 1399–1408, 2018.

[3] G. Cowburn and L. Stockley, “Consumer understanding and use of nutrition labelling: a systematic review,” *Public Health Nutr.*, vol. 8, no. 1, pp. 21–28, 2005.

[4] C. N. Mhurchu and D. Gorton, “Nutrition labels and claims in New Zealand and Australia: a review of use and understanding,” *Aust. N. Z. J. Public Health*, vol. 31, no. 2, pp. 105–112, 2007.

[5] D. Nelson, D. Graham, and L. Harnack, “An objective measure of nutrition facts panel usage and nutrient quality of food choice,” *J. Nutr. Educ. Behav.*, vol. 46, no. 6, pp. 589–594, 2014.

[6] J. Jewell, “What is the evidence on the policy specifications, development processes and effectiveness of existing front-of-pack food labelling policies in the WHO European Region?,” 2019.

[7] K. L. Hawley, C. A. Roberto, M. A. Bragg, P. J. Liu, M. B. Schwartz, and K. D. Brownell, “The science on front-of-package food labels,” *Public Health Nutr.*, vol. 16, no. 3, pp. 430–439, 2013.

[8] K. G. Grunert and J. M. Wills, “A review of European research on consumer response to nutrition information on food labels,” *J. Public Health (Bangkok).*, vol. 15, no. 5, pp. 385–399, 2007.

[9] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[10] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, 2004.

[11] W. Liu *et al.*, “Ssd: Single shot multibox detector,” in *European conference on computer vision*, 2016, pp. 21–37.

[12] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[13] M. Liao, B. Shi, and X. Bai, “Textboxes++: A single-shot oriented scene text detector,” *IEEE Trans. image Process.*, vol. 27, no. 8, pp. 3676–3690, 2018.

[14] Y. Liu and L. Jin, “Deep matching prior network: Toward tighter multi-oriented text detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1962–1969.

[15] M. Liao, Z. Zhu, B. Shi, G. Xia, and X. Bai, “Rotation-sensitive regression for oriented scene text detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5909–5918.

[16] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, “Detecting text in natural image with connectionist text proposal network,” in *European conference on computer vision*, 2016, pp. 56–72.

[17] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[18] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao, “Scene text detection via holistic, multi-channel prediction,” *arXiv Prepr. arXiv1606.09002*, 2016.

[19] D. He *et al.*, “Multi-scale FCN with cascaded instance aware segmentation for arbitrary oriented word spotting in the wild,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3519–3528.

[20] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, “Textsnake: A flexible representation for detecting text of arbitrary shapes,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 20–36.

[21] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, “Multi-digit number recognition from street view imagery using deep convolutional neural networks,” *arXiv Prepr. arXiv1312.6082*, 2013.

[22] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, “Synthetic data and artificial neural networks for natural scene text recognition,” *arXiv Prepr. arXiv1406.2227*, 2014.

[23] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] K. Cho *et al.*, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” *arXiv Prepr. arXiv1406.1078*, 2014.

[25] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, “Robust scene text recognition with automatic rectification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4168–4176.

[26] T. He, Z. Tian, W. Huang, C. Shen, Y. Qiao, and C. Sun, “An end-to-end textspotter with explicit alignment and attention,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5020–5029.

[27] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[28] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, “Real-time scene text detection with differentiable binarization,” in *Proceedings of the AAAI conference on artificial intelligence*, 2020, vol. 34, no. 07, pp. 11474–11481.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[30] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, “Multi-oriented and multi-lingual scene text detection with direct regression,” *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5406–5419, 2018.

[31] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, “Bag of tricks for image classification with convolutional neural networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 558–567.

[32] Y. He, P. Liu, Z. Wang, Z. Hu, and Y. Yang, “Filter pruning via geometric median for deep convolutional neural networks acceleration,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4340–4349.

[33] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, “Pruning filters for efficient convnets,” *arXiv Prepr. arXiv1608.08710*, 2016.