BLIDA 1 UNIVERSITY
Faculty of Sciences
Computer Science Department

Master's Thesis
In Computer Science

Option : Software Engineering

A Deep Learning Model For Food Pairing

Realised by
BENGOUFA Fady Ayoub

Supervised by
Dr. BACHA Siham

In front of the members of the jury

President :	Dr. CHERIGUENE	Blida 1
Examiner :	Ms. HADJ HENNI	Blida 1

September 2022

Abstract

Humans have come a long way, from nomads to farmers, from growing fruits and livestock farming for their survival to having the luxury of innovating with food that produced different types of cuisines and became the identity of different cultures.

Food science, multidisciplinary science that studies food's physical, biological and chemical aspects, has been around for centuries. One emerging aspect is the study of food pairing. Chefs have tested countless food ingredient pairs through trial and error and using their expertise in their respective cuisine styles, but this method is finite and consumes energy and resources.

In this study, we proposed two approaches based on deep learning techniques to create a model that predicts the scores of ingredient pairs. The first approach employs a Siamese Neural Network model that recommends ingredient pairs using the frequency of appearance of those pairs. The second approach focuses on recommending ingredient pairs that share similar flavor compounds. We have concluded that both models give us insights on how to innovate regarding pairing food ingredients. Where the first one considers familiar ingredient pairs, the second one recommends uncommon new pairs based on the food pairing hypothesis, which states that food with similar flavor compounds tastes good when consumed together.

Keywords: deep learning, food pairing, siamese neural network, food pairing hypothesis, natural language processing

الملخص

لقد قطع البشر شوطًا طويلاً ، من البدو إلى المزارعين ، من زراعة الفاكهة وتربية الماشية من أجل بقائهم إلى التمتع برفاهية الابتكار مع الطعام الذي ينتج أنواعًا مختلفة من المأكولات وأصبح هوية ثقافات مختلفة.

علم الغذاء ، علم متعدد التخصصات يدرس الجوانب الفيزيائية والبيولوجية والكيميائية للأغذية ، كان موجودًا منذ قرون. أحد الجوانب الناشئة هو دراسة اقتران الطعام. اختبر الطهاة عددًا لا يحصى من أزواج المكونات الغذائية من خلال التجربة والخطأ واستخدام خبراتهم في أنماط المطبخ الخاصة بهم ، لكن هذه الطريقة محدودة وتستهلك الطاقة والموارد.

في هذه الدراسة ، اقترحنا نهجين يعتمدان على تقنيات التعلم العميق لإنشاء نموذج يتنبأ بعشرات أزواج المكونات. يستخدم الأسلوب الأول نموذج شبكة سيامي العصبية الذي يوصي بأزواج المكونات باستخدام تكرار ظهور تلك الأزواج. يركز النهج الثاني على التوصية بأزواج المكونات التي تشترك في مركبات نكهة مماثلة. لقد توصلنا إلى أن كلا النموذجين يعطينا رؤى حول كيفية الابتكار فيما يتعلق بإقران المكونات الغذائية. عندما ينظر الأول في أزواج المكونات المألوفة ، يوصي الثاني بأزواج جديدة غير شائعة بناءً على فرضية الاقتران بالطعام ، والتي تنص على أن الطعام الذي يحتوي على مركبات نكهة مماثلة يكون طعمه جيدًا عند تناوله معًا.

الكلمات المفتاحية: التعلم العميق ، الاقتران الغذائي ، الشبكة العصبية السيامية ، فرضية الاقتران الغذائي ، معالجة اللغة الطبيعية

Acknowledgements

In the name of *Allah*, the most gracious and the most merciful. First and foremost, I am thankful to the Almighty *Allah* for giving me the strength, knowledge, ability and opportunity to under take this work and complete it satisfactorily. Writing this thesis definitely took a toll on me mentally but remembering that *surely with hardship comes ease* helped me keep a sane mind and commit to my work.

Second, my *Mother* who was the person who brought me to life and did her best to provide for me everything I needed. She was the person that bought me my first computer that allowed to learn and improve in this field, If it wasn't for that I wouldn't be where I am now.

I want express my deepest gratitude to my teacher *Mrs. Bacha* for her supervision, support and her patience for my stupid questions. If it wasn't for her help I wouldn't have been able to finish this thesis. I've met many teachers throughout school and university, very few left a good impression on me, and *Mrs. Bacha* is definitely one of the teachers that I admire and wish to be able to reach her level.

I am thankful for the friends that supported me and asked about me several times during this period.

Lastly I am thankful that this chapter of my life is coming to an end. Excited for what *Allah* has for me in the future. Hopefully I will leave a good trace before I depart from this world.

Dedication

This thesis is dedicated to my friends in *IT Community* my university club. This club allowed me to meet incredible individuals that increased my love and passion of computer science. It was a big part of university life that helped me survive these five years with a sane mind.

To mention some names of these individuals: *Tarek B., Wafa Nesrine B., Aimen B., Karim H., Abdelkarim N., Ayoub C., Faical B., Faycal B., Hamza H., Kenza S., Lydia A., Fatma Zohra M., Amira B., Billel T., Fethi G, Oulfa E., Loubna M., Merouane M., Lina O., Lydia Z., Sonia A., Yasser A., Abdou B., Nesrine B., Hiba O., Abdelaziz M., Meriem B.* and many more that I forgot their names.

I dedicate this thesis as well to my dear friends: *Yacine B., Meriem F., Nourelhouda M., Wassim M., Chemsddine B., Zyneb L.* and everyone who helped me fulfil this work directly or indirectly.

I dedicate this thesis to the people I look up to and appreciate their work they are doing in this world: *Yes Theory, Nathaniel Drew, Lex Fridman, Matt D'Avella, Anas Bukhash* and many more.

Allah Blessed me with meeting and befriending remarkable people and I will forever be grateful and thankful for it.

Table of Contents

List of Tables

List of Figures

General Introduction

”A recipe has no soul. You, as the cook, must bring soul to the recipe.” said the American chef and cookbook writer Thomas Keller. Food to some is their love language. Mothers put everything into cooking for their families, and it brings joy to them to see their children satisfied with their daily nutrition setting them up for their day. Mothers are considered heroes just for this act alone. Likewise, chefs and restaurant owners do not want their customers to sit and eat; they want them to stop eating and feel sick with longing. Chefs aspire to get Michelin stars, the highest award for outstanding cooking that considers the ingredients' quality, the harmony of flavors, the mastery, and the chef's personality as expressed in their cuisine.

Throughout the centuries, cooking has been achieved through trial-and-error, a tedious process, that is for sure, but that is the only option chefs had. Later on, different techniques came out that revolutionized cooking and opened new doors for innovation. Molecular gastronomy is a discipline concerned with the physical and chemical transformations that occur during cooking. The latest one is the food pairing hypothesis famously applied by Heston Blumenthal, chef of The Fat Duck Restaurant and pioneer of multi-sensory cooking, food pairing, and flavor encapsulation. The principle of food pairing is that ingredients combine well with similar molecular compounds. Data scientists have taken this hypothesis and combined technology to test it and create solutions that will help chefs in their journey.

The food pairing hypothesis, although it can help chefs to combine new ingredient pairs and create new recipes, it would take an enormous time and effort to experiment with different ingredients manually without the use of technology. So, can we create a food pairing solution that can be a beckon for chefs? Allowing chefs to try different styles of ingredient pairs regardless of the region?

Our work seeks to use artificial intelligence to create an easy solution to recommend ingredient pairs based on their frequency of appearance and based on ingredient pairs' shared flavor profiles. We will use the subfield of natural language processing specialized in textual data, and its techniques will help us process and clean our datasets and make them ready to use in training on deep learning neural networks.

One of the major difficulty that we encountered was the lack of ingredients with flavor compounds, which led us to use only recipes that had all flavor compounds of its ingredients. Another obstacle is the small dataset for the traditional algerian recipe data that allowed only the second approach based on shared flavor profiles to work.

In this thesis, we start first and foremost by going through in detail the food pairing science, its related work, and our contribution to this field. Next, we will explain the basics of artificial intelligence, emphasizing deep learning techniques and how neural networks function. In the third chapter, we explain further in detail our two proposed approaches based on deep learning and go through all the data processing stages and prepare our data input for neural networks training. In the fourth and last chapter, we present our implementation of the two proposed approaches, evaluate them and verify their results with test cases, alongside presenting our user interface that will recommend ingredient pairs based on the two approaches.

Finally, we conclude with a general conclusion of our research and present our future perspectives.

Chapter 1

Food Pairing

1.1 Context

“Tell me what you eat, and I will tell you what you are,” said the French gastronome Jean Anthelme Brillat-Savari. This phrase is not just a clever one-liner but a profound insight into our relationship with food. Throughout history, we have achieved food preparation through trial and error. The consummation and preparation of food depend on many factors: cultural, historical, religious, geographical, and climatic, everything that defines a given society. The religious factor has a significant impact on what religious individuals eat. The monotheist religions such as Islam, Christianity, and Judaism have strict rules on what not to eat and drink, which has affected all of their traditional food [56].

Humans are addicted to food. As such, they have created countless food shows worldwide that consist of food competitions and cooking shows like Top Chef, MasterChef, and many more [55].

Everyone thinks that we taste food with our tongues, but it is not entirely the case. The hidden science behind great ingredient pairs in recipes is that 80% of our flavor experience is using the smell, which explains why tasteful ingredient pairs are the ones that form strong aromatic matches, whereas the rest of the 20% count for the taste and touch [60] as illustrated in Figure 1.1.

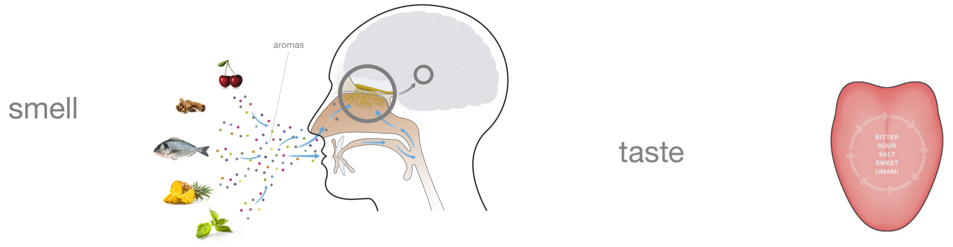


Figure 1.1: Aromas through smell and taste [60]

In the following pages of this chapter, we will go deep into explaining the food pairing science and its related work.

1.2 Food Pairing

Food is a crucial necessity for living creatures' survival. It has always been part of life so naturally. To satisfy their nutritional needs, humans throughout the years have found creative ways to create delicious food by combining random ingredients on just the basis of taste. Many factors influence food, such as culture and climate, but even with that, we have not fully explored the potential combinations of food in various regional cuisines [15].

Chefs, home cooks, and food engineers have used the food pairing method for decades. It is one of the most notable studies about food science, popularized by the famous chef of the Fat Duck restaurant and the author of the same book (The Big Fat Duck Cookbook) [19], Heston Blumenthal, in 1992 [16]. This method aims to create new possible food combinations to use in the culinary world. Many books tackle the question of improving the taste by food pairing [47][53] not only by instinct, cultural history, tradition, and plain guesswork but based on the food's aromatic molecules. Due to the importance of food pairing in food innovation, a food pairing startup[1]that helps chefs to find the most unusual combination based on science was created. It takes advantage of artificial intelligence to achieve this purpose. It allows access to a vast ingredients database to create the most fantastic novel combinations by matching ingredients with their best aromatic match. An example of food pairing experimentations is white chocolate with ciabatta (Italian white bread), gombo (green okra), emmental (cheese from Switzerland), and toast, thus pairing and assembling rare ingredient pairs with an innovative twist.

[1] ww.foodpairing.com

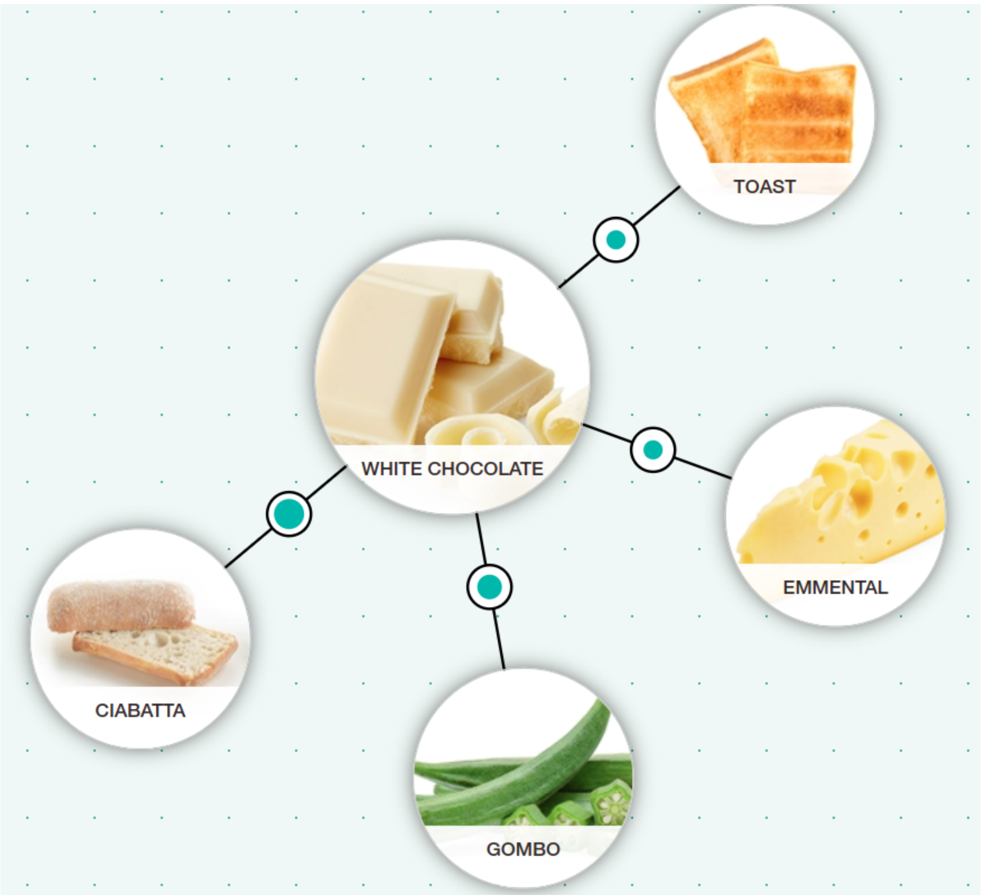


Figure 1.2: Combination of best matching ingredients out of white chocolate, created from the Food Pairing Platform[2]

For centuries, chefs and home cooks have been experimenting with different cooking styles to create new techniques and dishes used in famous restaurants. A movement began that has inspired chefs worldwide and will continue to inspire generations to come. One of the disciplines that helped this movement to start is the coming of Molecular Gastronomy [62].

[2] ww.foodpairing.com

1.3 Molecular Gastronomy

Molecular Gastronomy is a branch of food science that focuses on the physical and chemical methods that occur during the preparation and processing of food [30]. First was articulated in 1988 [50] by the french chemist Hervé [41] with the hungarian physicist Nicholas Kurti, and developed since then by my companies and research institutes.

Molecular gastronomy created numerous techniques to achieve the desired flavor or chemical reaction during food preparation. Popular ones include:

- **Foaming** is created by mixing liquids with emulsifiers [42] (combining two unmixable liquids to form a homogenous solution, used as a sauce or garnish on dishes)[30].
- **Carbonating:** Creating a bubbly effect in drinks by adding carbon dioxide, like carbonating sugar, creates air bubbles for a popping sensation in sweets and desserts [30].

There are many other methods to mention: Thickening, Glueing, Flash-freezing, etc. Some examples of food use molecular gastronomy to add flavor dimensions to meals.

- **Soufflés:** needs the right mixture of egg whites with the right ingredients, baked at the right temperature.[30]
- **Crème brûlée** crystallizes sugar using a flame torch and creating a crunchy coating [30].

Molecular gastronomy [62] combined with the culinary skills of the top chefs in the world resulted in the rise of their restaurants to be the best in the world. **ElBulli** is a spanish restaurant once led by the famous spanish chef Ferran Adrià. He revolutionized food by introducing techniques such as spherification [3] from fruit juices. His untraditional approach proved successful in manifesting techno-emotional dishes into reality. Figure 1.3 and 1.4 represents two famous restaurants, "El Bulli restaurant" and "El Celler de Can Roca."

Figure 1.3: El Bulli restaurant from above

”El Bulli is undoubtedly the most controversial and experimental restaurant in the world that receives up to 1,000,000 reservation requests a year, where only 8,000 lucky ones get a table.”[3]

Figure 1.4: El Celler de Can Roca restaurant, Girona, Spain

Another famous restaurant located in Girona, Spain, called El Celler de Can Roca[4]founded in 1986. Known for distinctive and unique food combinations of different geographical locations and climate conditions, from shrimps, lamb, duck, and mushrooms.

These were only some well-known restaurants that decided to take the traditional culinary techniques to the next level using molecular gastronomy, studying the chemical ingredients to create outstanding flavorful dishes and defy the status quo of the current culinary world.

In the next section, we will review the related work on the food pairing hypothesis.

1.4 Related Work

There have been few studies that tackled food pairing science. The most famous one is the food pairing hypothesis which states that ingredients sharing flavor compounds are more likely to taste well together than those that do not. However, this hypothesis is not valid for all regions. Asian cuisine argues the opposite, which makes it quite hard to find a pattern for creating food ingredient combinations [15][54].

Ahn et al. worked on a published scientific report on creating a flavor network for food pairing [15]. The results show that Western cuisines tend to use ingredient pairs that share many flavor compounds, but on the other hand, East Asian cuisines tend to share few flavor compounds. The food pairing hypothesis suggests that the more we use ingredients that have many chemical compounds in common, the more the food taste is better. This

[3] www.elbulli.info

[4] cellercanroca.com

hypothesis has been used to combine ingredients to create novel plates. However, the study goes against the food pairing hypothesis by building the flavor network, a bipartite network consisting of two types of nodes; the first node had 381 ingredients used in recipes from around the world. The second one had 1,021 flavor compounds contributing to the ingredients used on the first node. Each ingredient is connected with all their respective flavor compounds, as illustrated in Figure 1.5.A The weight of each link represents the number of shared compounds, as illustrated in Figure 1.5.B.

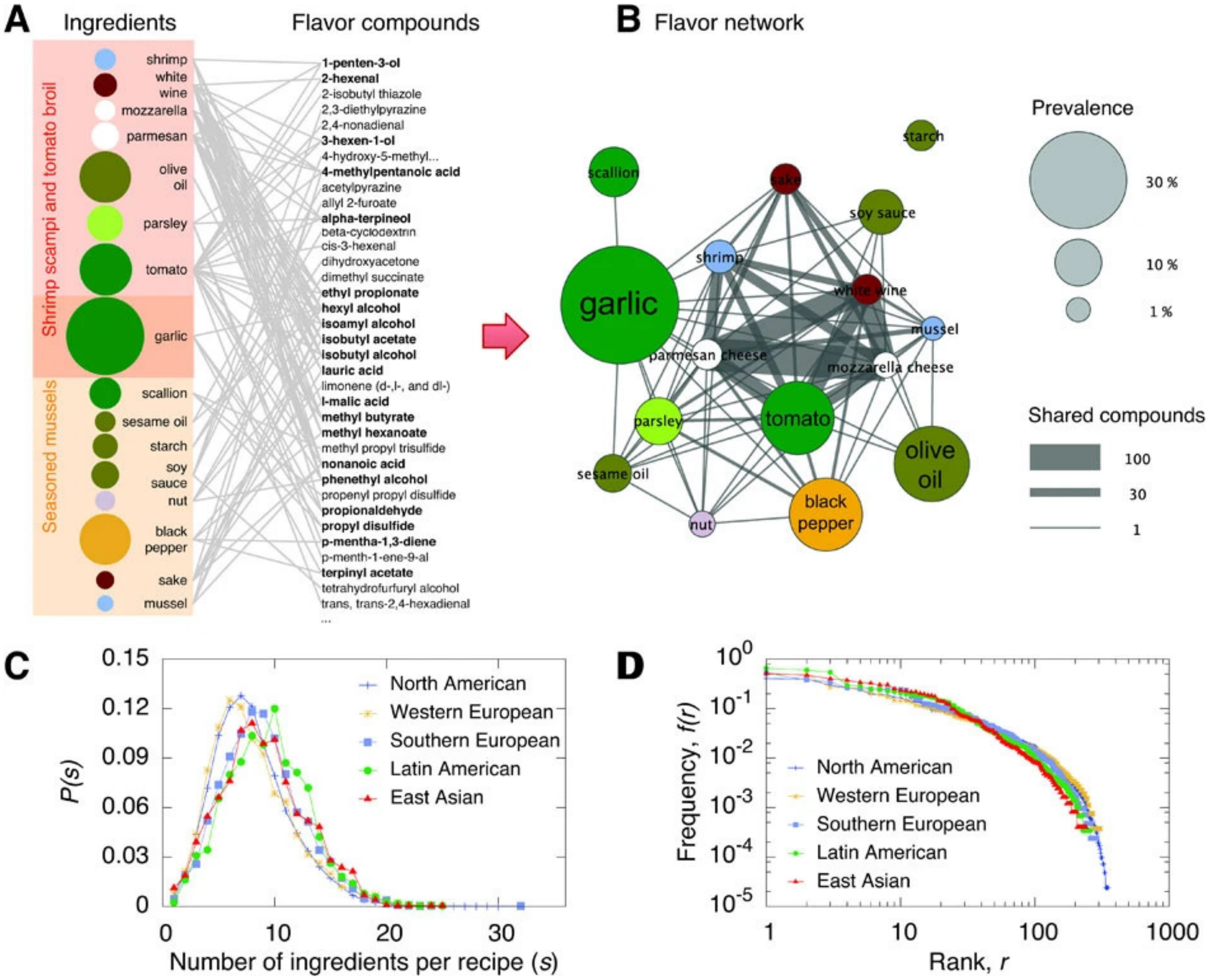


Figure 1.5: Bipartite Flavor Network [15]

Figure 1.5.D demonstrates that North American and Western European cuisines tend to have recipes where their ingredients share flavor compounds. On the contrary East Asian and Southern European cuisines avoid ingredients that share flavor compounds.

A considerable amount of ingredients share several flavor compounds, but The flavor network is too small to visualize all the ingredients. Therefore, they used an extraction method to identify links for each ingredient. Figure 1.6 shows the links between the different ingredients. The wider the link, the higher the flavor compounds shared.

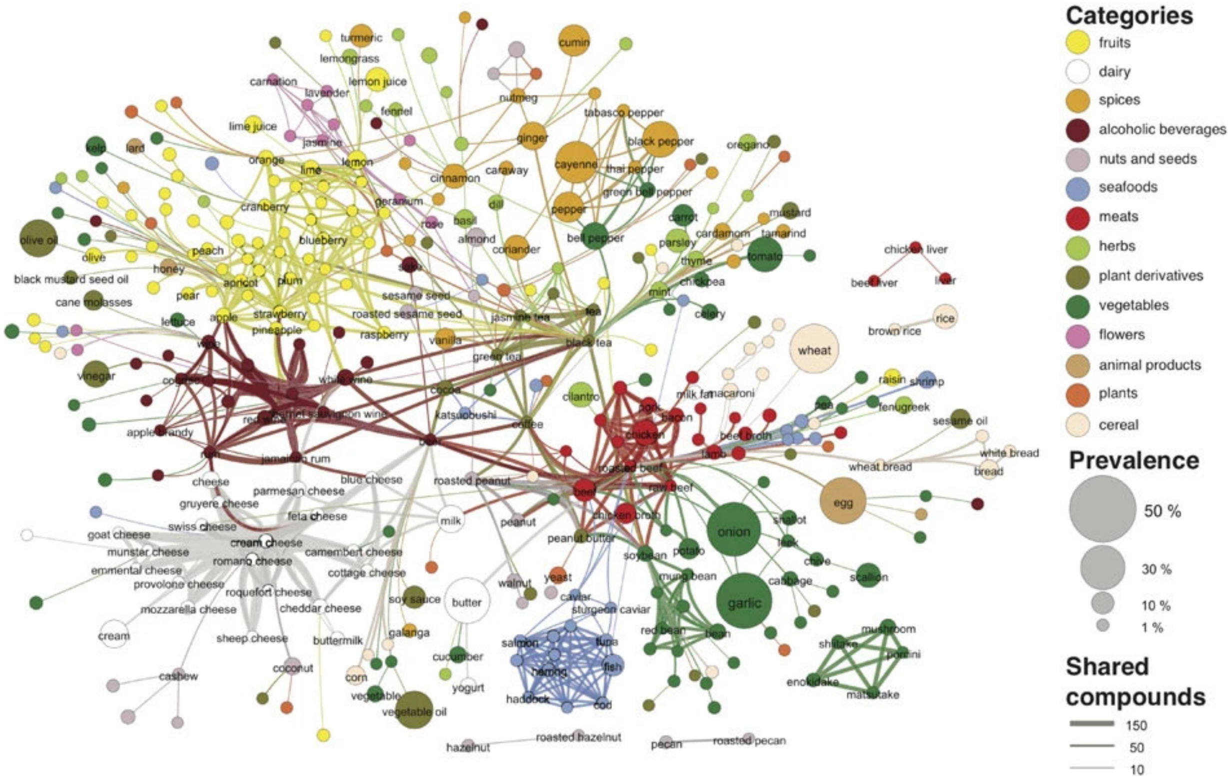


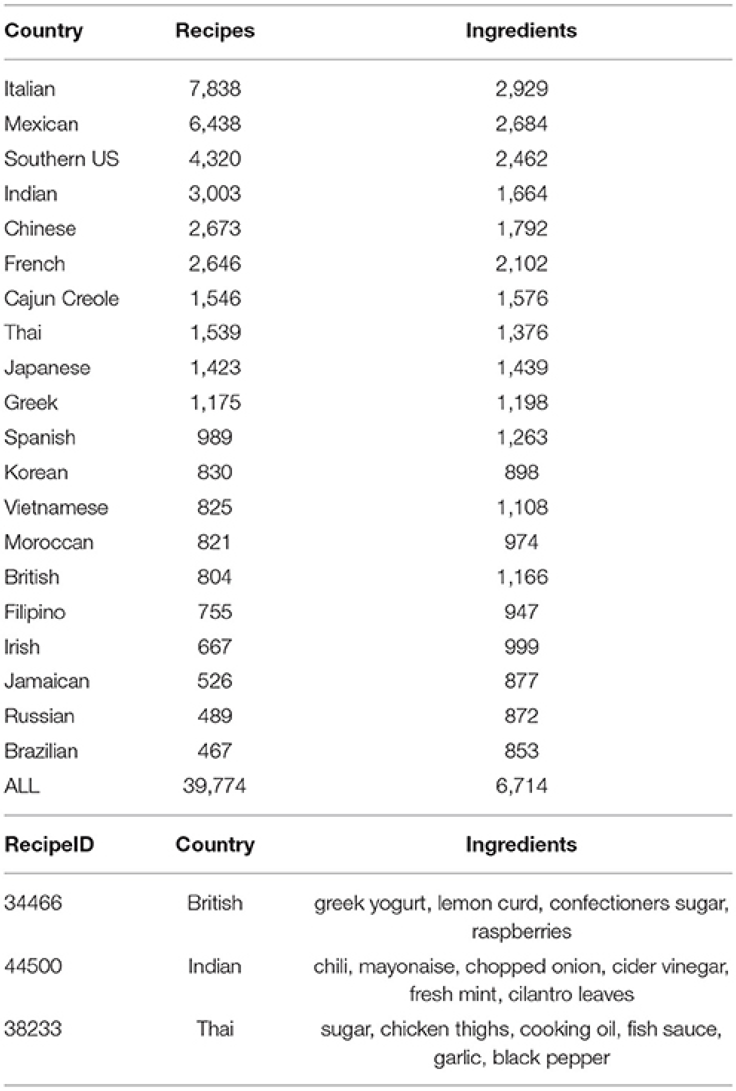
Figure 1.6: The Backbone of the flavor network [15]

Another study has been realized by Kazama et al. [35]. Given the enormous differences in recipes from different regions, this system can transform any recipe into any selected regional style (e.g., Japanese, Mediterranean, or Italian). It can also pinpoint the mixture of regional cuisines style of any selected recipe and visualize using the barycentric Newton diagrams.

The dataset used to train the deep learning model that detects the region a recipe originated from is the Yummly dataset [66] which has 39,774 recipes from 20 countries, as shown in Table 1.7.

In 2019 another study regarding food pairing was done by Park et al. [23]. It allows for predicting food ingredient pairing scores and recommends the pairs of the optimal ingredients; this model is called KitcheNette. KitcheNette used Siamese neural networks and was trained on a dataset called the Recipe1M+ [39] that contains approximately 1 million recipes and their corresponding images collected from multiple data sources found on cooking websites. 5% of the dataset are known pairings of ingredients with more than 300k of the possible number of ingredients pairs, where 95% total of unknown pairings of food ingredients that are either rarely or never used in recipes that they intend to predict their score with the KitcheNette model.

A recent study explores Saudi cuisine using genetic algorithms [16]; this is the first reported study in the Arab world. The study tested the food pairing hypothesis on Saudi cuisine, revealing a positive inclination similar to Western cuisine.



Country	Recipes	Ingredients
Italian	7,838	2,929
Mexican	6,438	2,684
Southern US	4,320	2,462
Indian	3,003	1,664
Chinese	2,673	1,792
French	2,646	2,102
Cajun Creole	1,546	1,576
Thai	1,539	1,376
Japanese	1,423	1,439
Greek	1,175	1,198
Spanish	989	1,263
Korean	830	898
Vietnamese	825	1,108
Moroccan	821	974
British	804	1,166
Filipino	755	947
Irish	667	999
Jamaican	526	877
Russian	489	872
Brazilian	467	853
ALL	39,774	6,714

RecipeID	Country	Ingredients
34466	British	greek yogurt, lemon curd, confectioners sugar, raspberries
44500	Indian	chili, mayonaise, chopped onion, cider vinegar, fresh mint, cilantro leaves
38233	Thai	sugar, chicken thighs, cooking oil, fish sauce, garlic, black pepper

Figure 1.7: Statistics of Yummly Dataset and some recipe examples [66]

In this research, we work on predicting the scores of food ingredient pairs and rank them based on the predicted score. We consider the molecule factor of the ingredients to predict novel food pairs with the ultimate taste.

1.5 Contribution

In this work, we are interested in Exploring matching ingredients that can taste well together. Table 1.1 below resumes all of the works mentioned previously.

Works	Problem Solved	Method Used	Datasets Used
Ahn et al. [15]	Testing the food pairing hypothesis	Flavor Network Graph	/
Kazama et al. [35]	Transform a recipe into any selected cuisine style.	Neural Networks	Yummly Dataset [66]
Park et al. [23]	Prediction of unknown ingredient pairs.	Siamese Neural Networks	Recipe1M+ [39]
Al-Razgan et al. [16]	Testing the food pairing hypothesis in Saudi Cuisine	Genetic Algorithm	Saudi and Middle-Eastern Cooking Book [16]

Table 1.1: The studies and experiments that were done in regard to the food pairing hypothesis.

Little research was done regarding the food pairing hypothesis to discover new ways to match unusual ingredients not typically used in the culinary world. In our work, we explore this hypothesis and look deep behind the science of molecular matchmaking of ingredients.

Furthermore, to contribute to our country's research, we will study the food pairing in traditional algerian cuisine. That makes our work the first country to study this hypothesis in the whole continent of Africa.

1.6 Conclusion

In this chapter, we introduced the food pairing hypothesis and how molecular gastronomy revolutionized the food industry, and we mentioned some studies done about food pairing. We will later see in the following chapters our contribution in detail.

Chapter 2

Artificial Intelligence

2.1 Introduction

In 1950 Alan Turing asked the question, “Can machines think?” In his paper titled Computing Machinery and Intelligence [61], he proposed a test called the Imitation game, aka the Turing Test, which tests the ability of a machine to exhibit intelligent behavior similar to a human being. However, unfortunately, Turing could not continue his work because the computers were not as powerful as they are now.

The term Artificial Intelligence was first presented in The Dartmouth Summer Research Project on Artificial Intelligence Conference by John McCarthy in 1956 [32]. This same conference is now a historical event that catalyzed the subsequent years of AI research.

Nowadays, artificial intelligence is Everywhere, whether we know it or not. Some examples are Voice Assistants like Siri and Alexa, Self-driving cars. It is even integrated into many social media platforms and viewing platforms like Youtube and Netflix that uses AI to determine the movies and videos that suit the user’s taste and preference.

2.2 Machine Learning

Machine Learning (ML) is a subfield of artificial intelligence (AI), first coined by Arthur Samuel [59], an American pioneer in the field of computer gaming and artificial intelligence. He defined *machine learning* as “the field of study that allows computers to learn without being explicitly programmed.“

The main focus of this field is acquiring historical data from the world and using it later on to create learning systems. ML is classified into three major categories: supervised, unsupervised, and reinforcement learning. We will go into further detail in the following sections.

2.2.1 Supervised learning

In supervised learning, we map an input to an output based on a dataset of input-output pairs called labeled data. Supervised learning uses patterns to predict the label values of unlabelled data [58].

"Applications in which the training data comprises examples of the input vectors, and their corresponding target vectors are known as supervised learning problems."[18]

For example, the computer can be "fed" data with images of dogs labeled as a dog and images of cats labeled as a cat. By training on the labeled data and finding patterns, the model will be able to identify unlabeled data and maps them to their correct identity, in our case, either a cat or a dog.

Supervised learning uses historical data to predict statistically possible future events. For example, it might be historical weather data to predict upcoming rainy or sunny days. Depending on the nature of the output, both classification and regression problems are supervised learning problems.

- **Classification:** Supervised learning problem that involves predicting a class label [20].
- **Regression:** Supervised learning problem that involves predicting a numerical label [20].

2.2.2 Unsupervised learning

In unsupervised learning, our data is unlabeled. It is up to the learning algorithm to find common hidden patterns in the input data. "In unsupervised learning, there is no instructor or teacher, and the algorithm must learn to make sense of the data without this guide."[31]

There are many techniques of unsupervised learning, and clustering is the most commonly used. It refers to grouping data points with similar characteristics and assigning them to clusters. A cluster is a collection of similar objects, as shown in Figure 2.1 below.

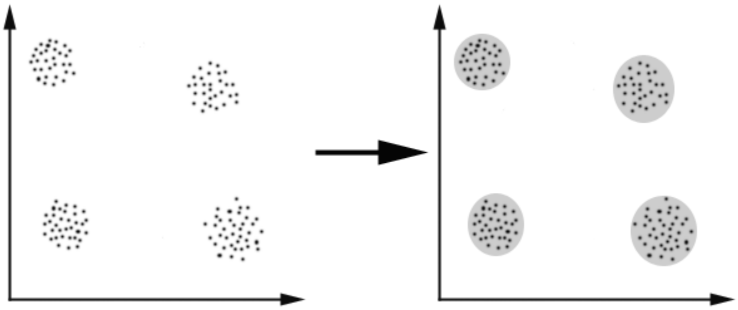


Figure 2.1: A collection of Clusters [43]

The goal of clustering is to determine the grouping of our unlabeled data. What decides a good clustering or not depends on the user defining the criterion so that the clustering solves the problem at hand.

2.2.3 Reinforcement learning

“Reinforcement learning is learning what to do — how to map situations to actions—so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them.”[21]

Reinforcement learning is where an agent operates in an environment and must learn to operate using feedback, meaning the system will be rewarded for desired actions and punished for undesired ones. Reinforcement learning (RL) is implemented in many fields, such as robotics, healthcare, and gaming.

One of the most famous computer programs, AlphaGo, developed by **DeepMind Technologies** [10]; used game optimization through RL and successfully was able to defeat the strongest Go (a famous strategic game)[1]player in the world in October 2015 and thus becoming the strongest Go player. The way Alpha Go trained was by starting to play amateur games in order to understand how a human plays. Then they had it play with itself thousands of times, learning from its mistakes each time. Over time it became better and better, thus becoming arguably the strongest Go player in history. Such a process is called reinforcement learning.

2.2.4 Deep learning

Deep Learning is a subfield of machine learning based on artificial neural networks, which function similarly to the neurons in our brains. Neurons are the building block of how our brain operates. Each neuron is designed to transmit information using electrical impulses and chemical signals between different areas in our brain and the rest of the nervous system. Deep Learning is made to mimic the same behavior by having deep stacks of layers of neurons. This depth of layers enables us to create deep learning models that handle complex hierarchical patterns found in challenging real-world data.

A simple Artificial Neural Network (ANN) is composed of node layers that have some neurons with an input value (Input Layer) and some output value (Output Layer). All are interconnected with another layer (Hidden Layer) or multiple layers.

The difference between an Artificial Neural Network (ANN) and a Deep Neural Network (DNN) is that an ANN can be shallow or deep. They are called shallow because they only have one hidden layer between the input and output layers. Where the DNN has more than one hidden layer, this is where the expression DNN (Deep Neural Network) comes from. Another thing is that an ANN can only have fully connected layers where all the inputs from one layer are connected to every neuron of the next layer. On the other hand,

[1] Go board game, ww.online-go.com,

a deep learning network can be much more complex and not have all of its layers fully connected layers.

The figure 2.2 down below is a fully connected deep neural network with five dense layers that include one input layer and one output layer with three hidden layers.

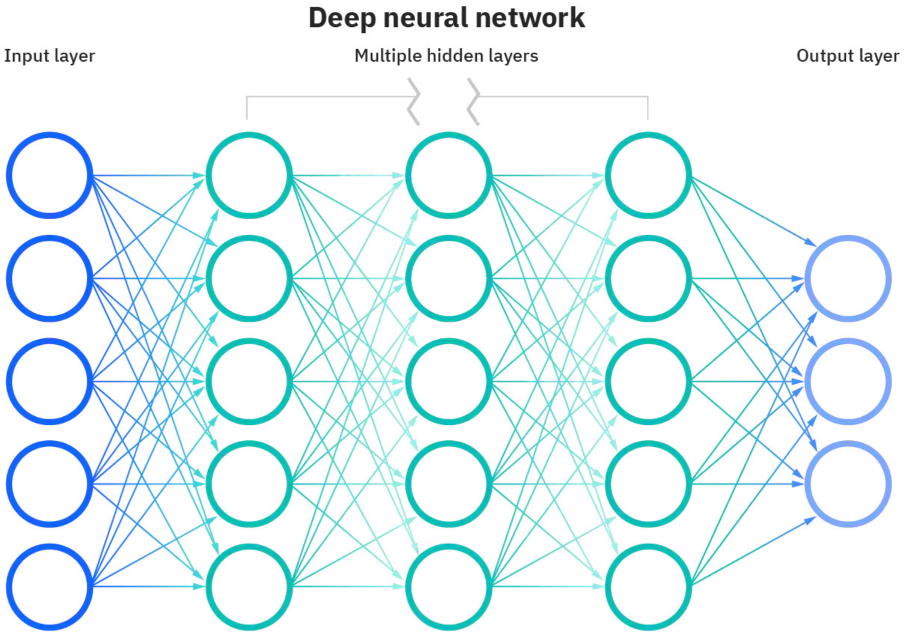


Figure 2.2: A Deep Neural Network [25]

Take the example of a dataset of house prices according to the number of rooms and bathrooms it has and the dimensions of houses. We can create a deep learning network algorithm that can predict the price of new houses.

The algorithm will try to minimize the difference between the predicted and expected outputs; this is how our model trains. It will learn the patterns and associations between the inputs (house characteristics) and the outputs (house prices). Thus after learning for the labeled dataset, our model can predict the unlabeled dataset because it is already a trained real-world dataset.

Let us first go through some basics to understand how that model trains and learns those patterns.

2.2.5 The Building Block of Neural Networks

Let's first understand the Linear Unit $y = a * x + b$ of one individual neuron in Figure 2.3

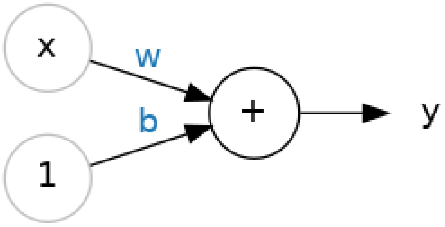


Figure 2.3: The Linear Unit: $y = w * x + b$[2]

The input x is connected to the neuron with a weight w, and the connection is then $w * x$. The b is also considered a weight called the bias; it is not connected to the input x, which is why in the example, its value is 1. The bias allows us to activate the neuron independently without modifying the input. The y is the output.

So we say the formula for the neuron's activation is: $y = w*x+b$ it is also the equation of a line.

Layers of neurons organize every neural network; collecting all of the linear units of each neuron, we get a Dense Layer as shown in Figure 2.4.

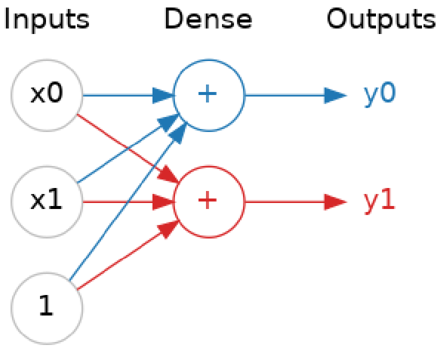


Figure 2.4: A dense layer of two linear units receiving two inputs and a bias[3]

Neural networks on their own can only learn linear relationships. It cannot learn curved relationships. That is why we need activation functions to fit a nonlinear relationship to its data points, as shown in Figure 2.5.

[2] www.kaggle.com/code/ryanholbrook/a-single-neuron
[3] www.kaggle.com/code/ryanholbrook/deep-neural-networks

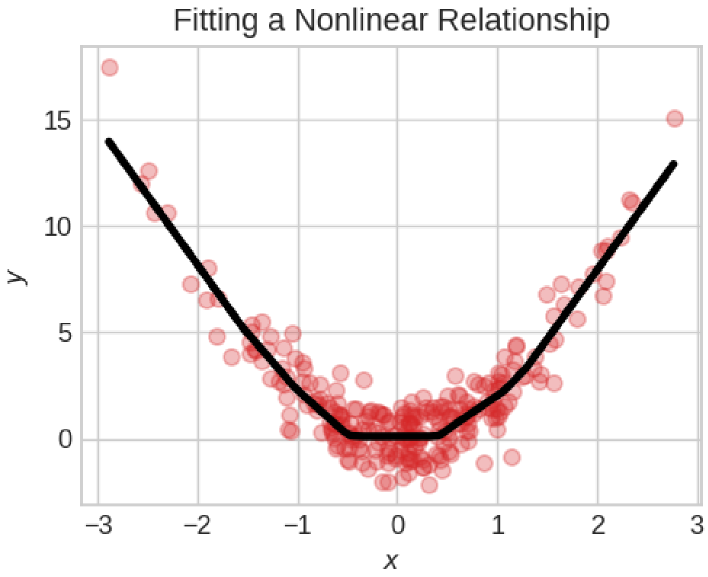


Figure 2.5: Fitting a Nonlinear function to its data points[4]

2.2.6 Activation Functions

Every neuron in our brain tries to segregate between valuable and non-useful information. This process is achieved in the artificial neural network using activation functions to ensure that the network only learns helpful information.

The activation function decides if the neuron should be activated (fired) or not using some mathematical operations. It will transform the sum of all the weighted inputs from the neuron into an output value that will be fed into the next layer [17]. a demonstration is shown in Figure 2.6 below.

[4] www.kaggle.com/code/ryanholbrook/deep-neural-networks

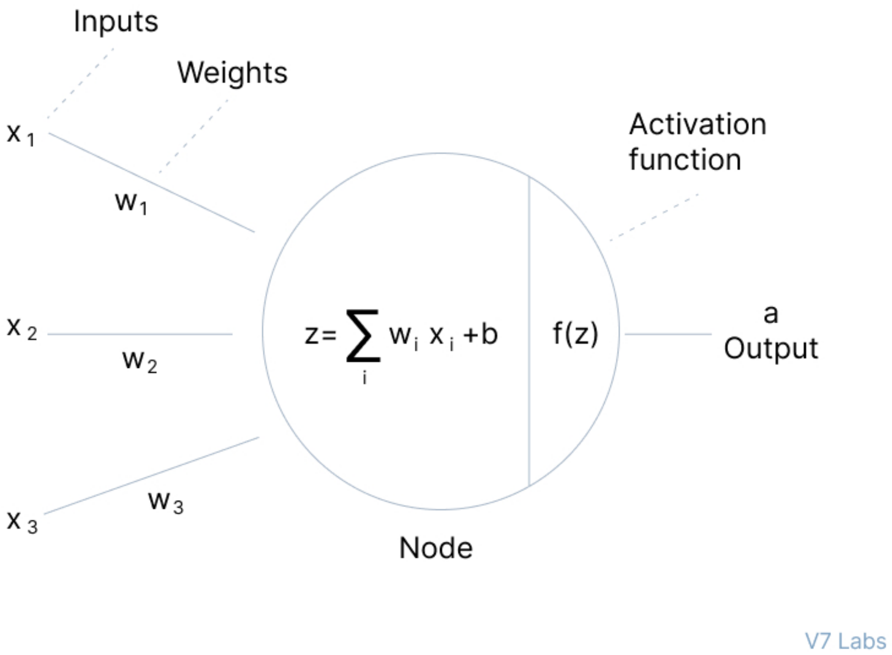


Figure 2.6: How an Activation Function work [17]

The most common activation function is rectifier function $max(0, x)$. The graph of the function is show in figure 2.7.

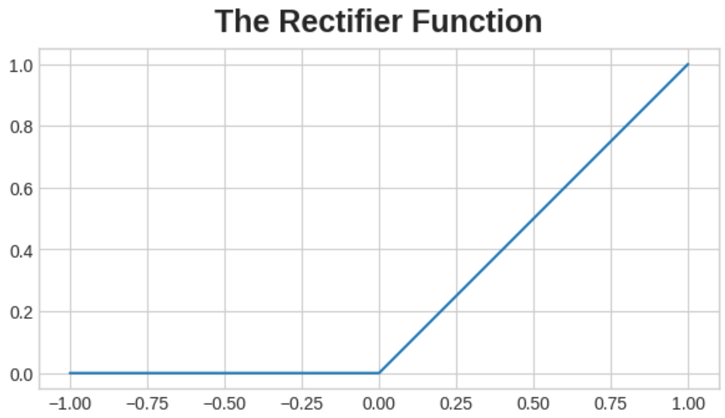


Figure 2.7: *relu* Activation Function [17]

Applying the *relu* activation function will move us out of the linear relationship by bending our data into a non-linear curve.

There are other activation functions such as *Sigmoid and Softmax*, the functions are chosen based on what kind of problem we are trying to solve and the values we went to get.[22]

2.2.7 Neural Network Training

Well, at first, the neural network does not know anything. The weights and biases of all of the neurons are initialized randomly. The idea is to correctly adjust the weights and biases to represent the inputs and outputs expressed in the training data. This allows us to have a trained model to predict any new data.

When learning, two movements are happening inside the neural network:*Feedforward Propagation and Backpropagation.*

Feedforward Propagation *(forward pass)* is the flow of information in a forward direction from the input layer (left) to the output layer (right). The input is used to calculate some intermediate function in hidden layers, later used to calculate an output. During this process, the activation functions are considered Gates between the input feeding the current neuron and the output going to the next layer.

Backpropagation is a process in which we adjust the weights and biases to minimize the error of the predicted output vector and the desired output vector. This computed error is called the loss function.

The loss function measures the disparity between the target's actual value and the model's prediction value.

Depending on the problem at hand, there are many loss functions to use. For a regression problem meaning the desired output are numerical values we use a common loss function called the Mean Absolute Error or MAE. for each training sample, we measure the MAE for the true value y_true and the predicted value y_pred by absolute difference $|y\_true - y\_pred|$ as shown in Figure 2.8.

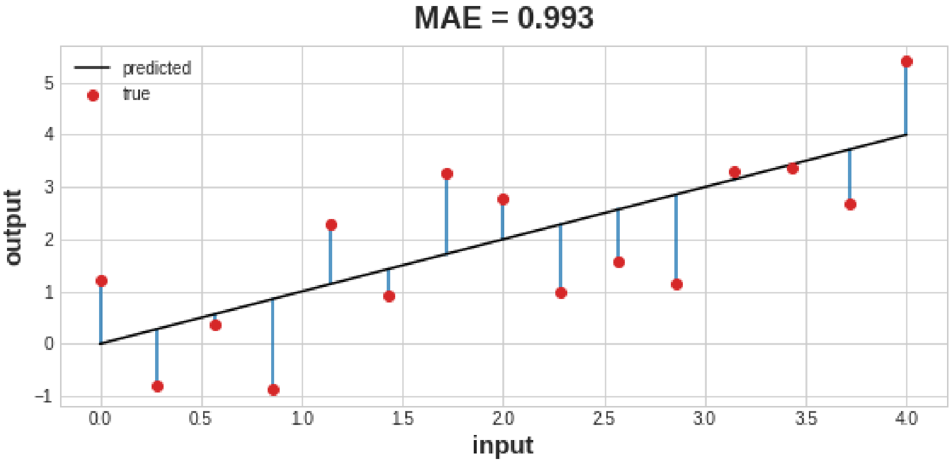


Figure 2.8: The mean absolute error is the average length between the fitted curve and the data points[5]

[5] www.kaggle.com/code/ryanholbrook/stochastic-gradient-descent

Other loss functions to mention for regression problems include the Mean Squared Error (MSE) or Cross Entropy loss, a multi-class loss.

The neural network, while training, will use the loss function as a guide to correctly adjust the weights and biases. The lower the loss function, the better the model's predictions are.

The loss function measures how good our predictions are in one training simple. In contrast, the cost function measures the average value of all training samples.

Therefore we try to optimize the cost function rather than just the loss function by using a specific optimizing algorithm.

2.2.8 Optimization algorithm of Gradient Decent

The optimizer algorithms are used to adjust the weights and biases by having the loss function as its guide. The most famous one in machine learning is the Gradient Decent Algorithm.

Gradient descent is an iterative optimization algorithm for finding the local minimum of a function. In our machine learning problem, that function is the cost function, as illustrated in Figure 2.9.

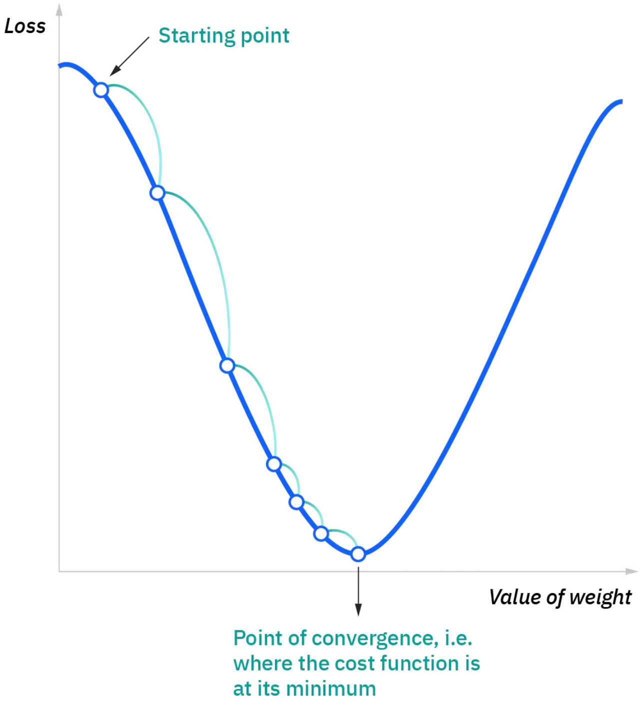


Figure 2.9: Finding Local Minimum using Gradient Decent [25]

2.2.9 Types of Neural Network Architectures

There are many types of neural networks used in deep learning problems; we mention three: Artificial Neural Networks (ANN), Convolution Neural Networks (CNN), and Recurrent Neural Networks (RNN). Each is used specialized in specific problems. For example, CNN's are more suitable for image problems, such as Object Detection Models or Image Classification.

In this thesis, the architecture that is used to solve our problem is the Siamese Neural Network.

2.3 Siamese Neural Network

Siamese Neural Network (SNN) is an architecture that contains two identical subnetworks, aka twin networks. Each sub-network has the same parameters, the same number of layers, and the same architecture. They also share the weights. They work in parallel by creating vector representations for the inputs and comparing their outputs by the end.

Let us take, for example, an Image Classification problem with CNN Architecture for our twin networks shown in Figure 2.10.

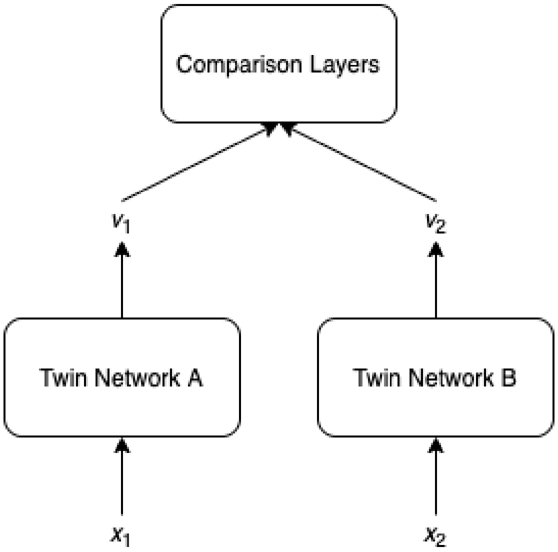


Figure 2.10: Siamese Neural Network [65]

In the graph above, $x1$ and $x2$ are two inputs we want to compare. $v1$ and $v2$ are their vector representation. The comparison layers architecture depends on the loss function and labels we want to train.

The idea here, given each pair of $(x1, x2)$ fed into each of the sub-networks, outputting a pair of vector representations $(v1, v2)$, we want the comparison layers to compute a similarity function depending on the final goal *(Euclidean Distance, Cosine Similarity and others).*

Figure 2.11 is a basic neural network architecture. Where Figure 2.12 shows a basic SNN architecture.

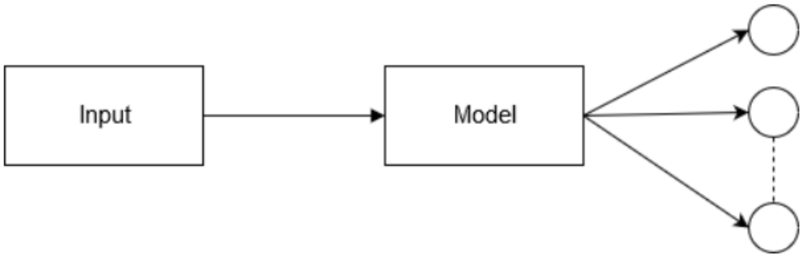


Figure 2.11: a Basic Neural Network[6]

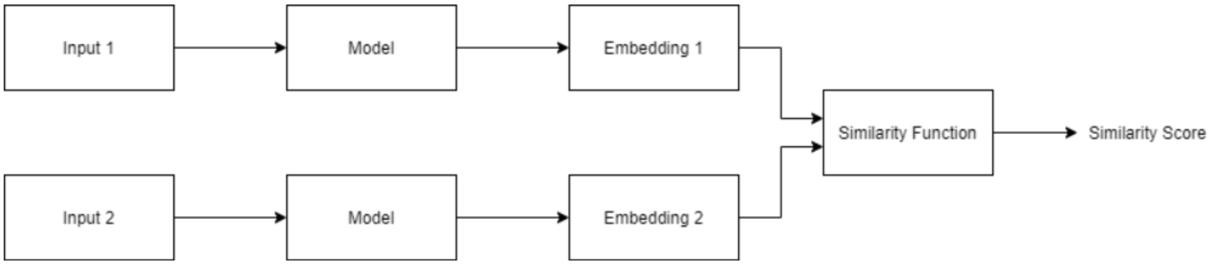


Figure 2.12: a Detailed Siamese Neural Network[7]

2.4 Natural Language Processing (NLP)

Natural language processing (NLP) is another branch of artificial intelligence concerned with understanding human spoken and written language.

The field of NLP has existed for more than 50 years and is rooted in linguistics. Medical research, search engines, and business intelligence are among the fields where it is used in real-world settings [38].

Human language is so complex and filled with ambiguities, making it difficult to create software to decipher their intended meanings. For a machine to understand text, several processes need to be applied to create valuable applications, such as Speech recognition or Sentiment Analysis [24].

In order to create an end-to-end NLP Software that can be used on new samples, we must go through certain stages; this is called the NLP Pipeline. It is not a universal set of steps, but it depends on the problem we are trying to solve.

[6] www.kaggle.com/code/sauravjoshi23/text-classification-using-siamesenet-glove/notebook

[7] www.kaggle.com/code/sauravjoshi23/text-classification-using-siamesenet-glove/notebook

- **Data Collection (Acquisition):** The first step in any NLP Task is to acquire a large volume of textual data.
- **Data Processing (Cleaning):** For the data to be used in the model, it needs to be cleaned and put in a way that highlights its features for the model to learn properly. This process, alongside the data collection, is the hardest and most time-consuming process of NLP Tasks.
  - **Text Cleaning:** Cleaning our text from HTML Tags, punctuation, and non-useful characters.
  - **Tokenisation:** Splitting our text into smaller units to work with it.
  - **Stopword Removal:** remove common words, or remove words that have no additional information that would help us such as words like: as, and, the, etc.
  - **Lemmatization and Stemming:** Reduce similar words to their common root, such as: "like" and "liked," which are the same work in different tenses[37].
- **Feature Engineering:** is the process of converting text data into numerical data. Because our machine learning model only understands numbers. One of the most used ways for that is:
  - **Bag of Words (BoW):** It is a representation of text that describes the occurrence of words within a document, disregarding the order of words and their structure.
  - **TF-IDF (term frequency-inverse document frequency) :** is a statistical measure that evaluates how relevant a word is to a document in a collection of documents.[57]
  - **Word Embeddings:** a type of word representation that allows words with similar meanings to have a similar representation. each word is represented as a real-valued vector. The vector's values are learned similarly to how a neural network learns.
- **Model Training:** it is the last phase is picking a model that can solve the NLP task at hand using the cleaned data and later the model would be tested on real data to predict new samples.

2.5 Conclusion

In this chapter, we provided an overview of Artificial Intelligence. We presented deep learning and machine learning fields l that can be used in a real-world problem, alongside the different techniques used to create such a model.

Chapter 3

Proposed Approach

3.1 Introduction

In this chapter, we will present the proposed approaches for predicting ingredient pairs that match together in a recipe. For that, we have proposed two approaches based on deep learning. The first one calculates the correlation between ingredients to identify if there are a known pairing or unknown pairing and their relative scores. Then, Siamese Neural Network is used to classify the ingredient pairs after extracting their embedding representation. The second approach creates a molecular profile for each ingredient, which will be employed by a deep learning model, alongside the correlation, to predict food pairing.

3.2 Global Scheme of the first approach

Figure 3.1 below shows a global overview of the first proposed approach based on a Siamese Neural Network that predicts Unknown Ingredient Food Pairing. The proposed approach contains five stages: Data collection, data processing, feature extraction using word embeddings, calculation of the correlation of ingredient pairs, and lastly, the training of our SNN Model to predict Unknown Pairings.

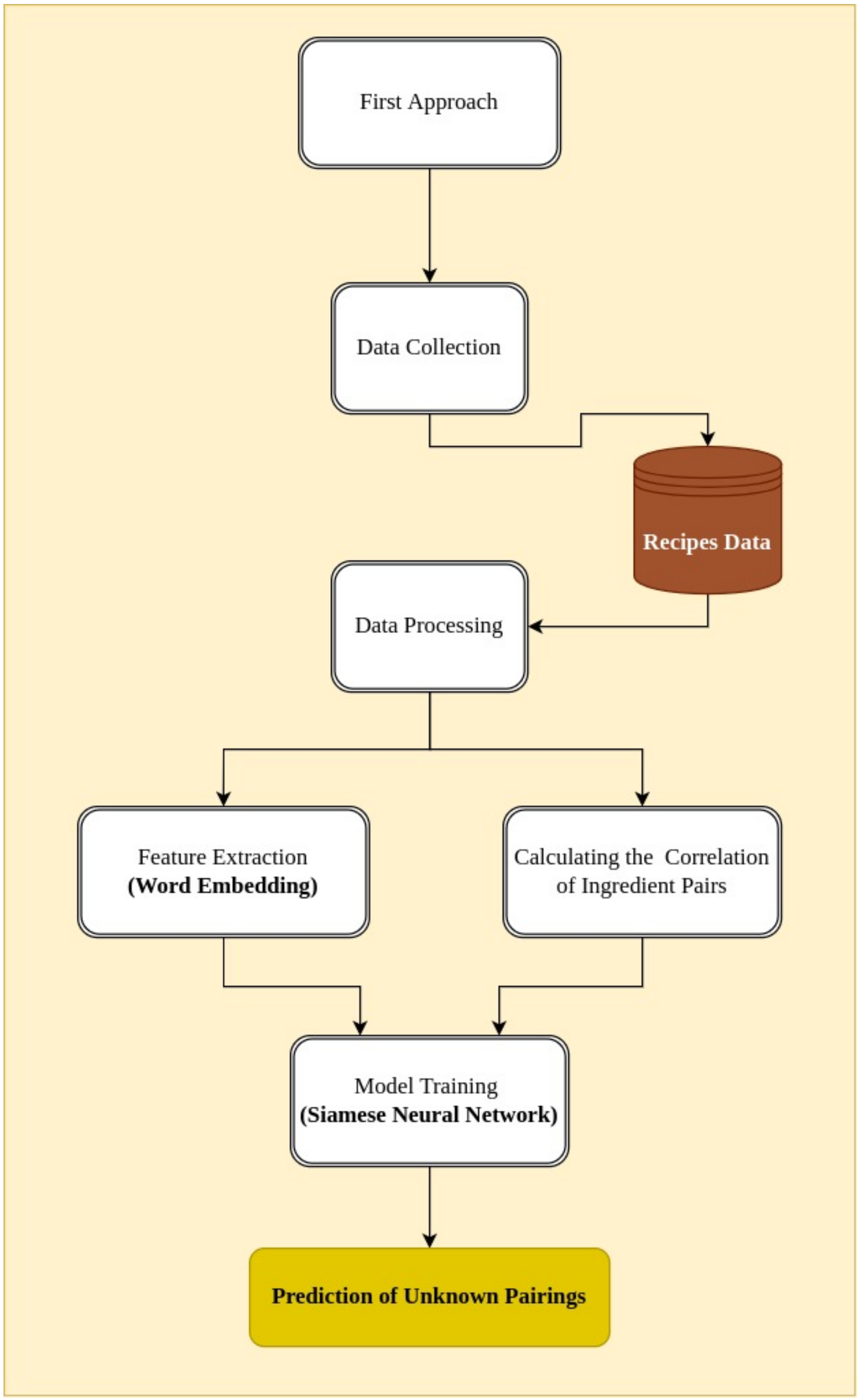


Figure 3.1: An overview of the second proposed approach based on the Deep Neural Network In the following, we will present each step of the second approach in detail.

3.3 First approach

3.3.1 Data Collection

To study the food pairing problem, we need a corpus of recipes dataset. We used the simplified-recipes-1M Dataset [52] by Dominik Schmidt, which contains over 1 million recipes.

3.3.2 Data Processing

The simplified-recipes-1M Dataset was already processed, But the data was not in the proper format; we needed to clean it so it could be ready for use by the proposed models.

Data Cleaning: this step includes removing all the errors in the dataset. In our recipes, data showed some problems, such as ingredients that were considered as ingredients but were not so, for example, ingredients like *yellow, dry, prepared, other.*

Moreover, some redundant ingredients were in singular and plural forms; for example, *potatoes and potato* are the same, so we need to keep only the singular ingredients in all of the recipes.

3.3.3 Feature extraction: Word Embedding

To train a deep learning model that trains on ingredient pairs, we need to use word embedding as input vectors to train our neural network.

Word embedding is a feature engineering technique used so that words or phrases for our vocabulary are mapped into vectors of real numbers.

For example, let us represent the two words "dad" and "mom" with a vector of real numbers length of 3:

$$\begin{aligned} dad &= [0.1548, 0.4848, 1.864] \\ mom &= [0.8785, 0.8974, 2.794] \end{aligned}$$

Furthermore, word embedding makes similar words in a semantic sense have similar vector representations. Words such as "dad" and "mom" are closer mathematically than words like "mom" and "cat" or "dad" and "butter."

For efficient training, we will represent our ingredients into a larger dimensional vector instead of length 3 in the previous example.

3.3.4 Correlation

In our learning model, we used the correlation statistical correlation measure to represent the relation between two pairs of ingredients and whether they work together well or not. Let us first understand what a correlation is.

Correlation is a statistical measure that expresses the strength of an association between two variables.

The correlation value is referred to as the correlation coefficient denoted by r. It ranges from $[-1, 1]$. The closer the value to 1, the stronger the relationship would be. When one variable increases, the other increases; this is called a positive correlation. When one increases, the other decreases; this is a negative correlation [14] [12].

The correlation used in our work is the Pearson's correlation coefficient. It is usually represented by $p(rho)$ 3.1.

$$p(X, Y) = cov(X, Y)/\sigma X \cdot \sigma Y \tag{3.1}$$

cov is the covariance. σX is the standard deviation of X and σY is the standard deviation of Y. The given equation for correlation coefficient can be expressed in terms of means and expectations 3.2.

$$p(X, Y) = E \cdot (X - \mu x) \cdot (Y - \mu y)/\sigma X \cdot \sigma Y \tag{3.2}$$

μx and μy are mean of x and mean of y respectively. E is the expectation[29] [28].

To implement the correlation on our recipe data, we need to create a binary matrix for our recipe dataset, as shown in Table 3.1 below.

# Recipes	**Ingredient 1**	**Ingredient 2**	...	**Ingredient N**
Recipe 1	1	0	...	1
Recipe 2	0	1	...	1
...	...	...	...	...
Recipe M	1	1	...	0

Table 3.1: Binary Matrix of Recipes Dataset

As shown in the table above, each row represents one recipe, and the columns are all of the unique ingredients used throughout the recipes. If the ingredient exists in the recipe, its value is 1; otherwise, it is 0.

It will make it easy to directly apply the correlation function that calculates the pairwise value for all the columns (ingredients).

The results are shown in Table 3.2 below.

# Ingredients	meat	butter	...	salt
meat	1.000000	0.204054	...	0.185020
butter	0.204054	1.000000	...	0.119111
...	...	...	...	...
salt	0.185020	0.119111	...	1.000000

Table 3.2: Correlation Matrix of Ingredient Pairs

Now that we have all the vector representations of all ingredients and the correlation of all pairs of ingredients. We need to feed it as input to our Model Classifier to be trained.

3.3.5 Siamese Neural Network Architecture

The Training Model we will use is the Siamese Neural Network architecture. We still need to restructure the data to feed our SNN model. For this purpose, data inputs constitute all possible combinations of ingredients which is the cartesian product of all unique ingredients. The ingredients will be replaced by their word embedding vector instead of their string value, finally, adding the correlation value of the pair as their label as shown in Table 3.3.

Ingredient	**Vector Ingredient 1**	**Ingredient 2**	**Vector Ingredient 2**	**Correlation**
milk	[1.6321754, 0.3437606, 0.50175226. . .]	sugar	[0.46397656, 0.51027924, 0.60459244. . .]	0.169771
milk	[1.6321754, 0.3437606, 0.50175226. . .]	vanilla	[0.22312754, 1.1426834, 0.6840022. . .]	0.124847
...	...	...	...	...

Table 3.3: Our input data for our SNN Model

We will consider all positive correlations of ingredient pairs as known pairs and the negative correlations as *unknown pairs* that we will predict later. The classification task consists of predicting if ingredient pairs work well together or not. For each sub-network, we used one Dense layer with *relu* as an activation function and an output layer of two nodes.

$$Dense(32, activation = relu)$$
$$Dense(2, activation = None)$$

For the Similarly Function, we computed manhattan distance, as shown in Figure 3.2 for the outputs of each sub-network to get a prediction value of range $[0, 1]$. Depending on the threshold we set during the training, we will have either a successful pair of 1 or an unsuccessful pair of 0. For that, we need to have a *threshold* that separates the ingredient pairs with a label of 1 (correlated ingredient pairs) and the ingredient pairs with a label of 0 (noncorrelated ingredient pairs).

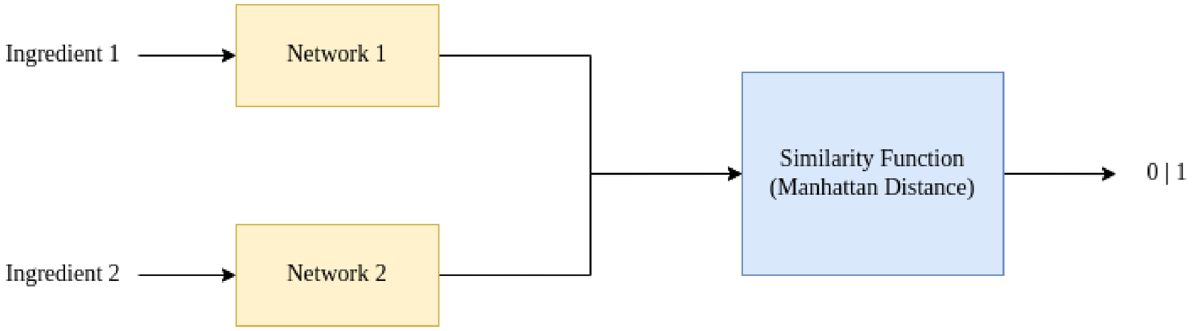


Figure 3.2: First Approach Siamese Network Architecture

3.4 Global Scheme of the second approach

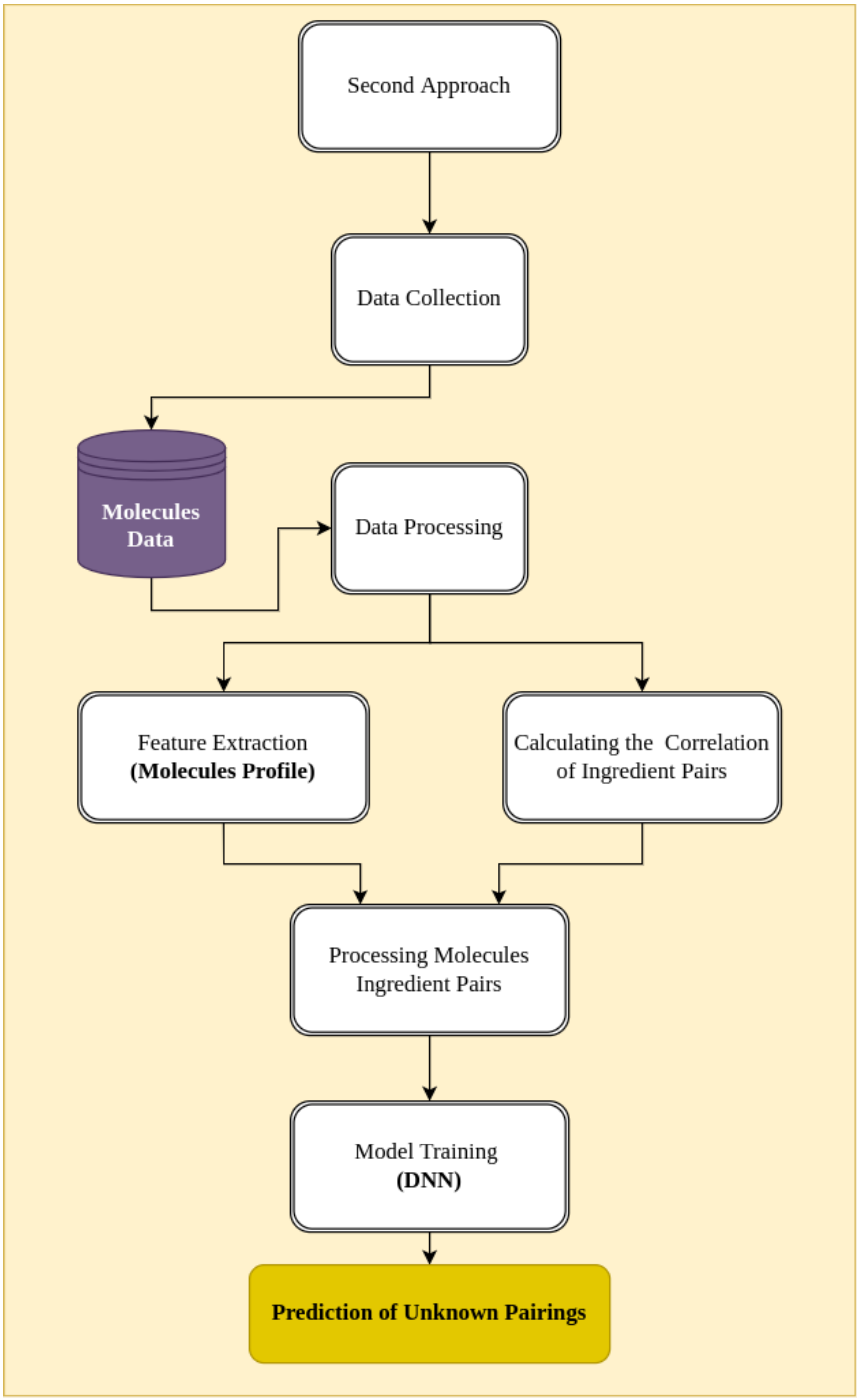

Second Approach
Data Collection
Molecules Data
Data Processing
Feature Extraction (Molecules Profile)
Calculating the Correlation of Ingredient Pairs
Processing Molecules Ingredient Pairs
Model Training (DNN)
Prediction of Unknown Pairings

Figure 3.3: An overview of the second proposed approach based on the Deep Neural Network In the following, we will present each step of the second approach in detail.

3.5 Second proposed approach

3.5.1 Overview of the second approach

To further explore the food pairing problem, a second approach was proposed. The first one takes the correlation and word embeddings only. In the second approach, we add the intersection of molecules for every ingredient pair as features and see what can be seen as a result.

3.5.2 Dataset

For the second approach, we need to create the molecular profile of each ingredient. For this, we use a dataset of Molecules named FlavorDB Dataset that contains many ingredients and their respective flavor molecules.

3.5.3 Data Cleansing

For our second approach to work, each ingredient needs to have its respective flavor compounds (molecules) existing in the FlavorDB dataset. However, unfortunately, we found many ingredients that did not exist in the FlavorDB, so we needed to remove the recipes whose ingredients did not have their molecules. Also, some ingredients needed to be changed, as illustrated in Table 3.4.

Ingredient Name	**Ingredient that exists in the FlavorDB**
Hot water	Water
Cheddar	Cheddar Cheese
Light cream	Cream
Apple Juice, Orange Juice, Lime Juice	Fruit Juice

Table 3.4: Replaceable Ingredients from FlavorDB

The ingredients that we could not replace for reasons we considered the whole recipe useless and removed them.

3.5.4 Feature extraction: Ingredient's Molecule profile

For this part, we need to create a dataset of molecules. We use the FlavorDB existing dataset that has over 900+ ingredients with their flavor compounds.

Each molecule has a chemical identifier, a common name, and a flavor profile, as shown in Table 3.5 below.

	Pubchem Id	**common name**	**flavor profile**
0	4	1-Aminopropan-2-ol	fishy
1	49	3-Methyl-2-oxobutanoic acid	fruity
...	...	...	...
1788	10340	Sodium Carbonate	odorless
1789	24856	Potassium alum	odorless
1790	24403	Tetrasodium Pyrophosphate	odorless

Table 3.5: Molecules and their chemical Id, common name and flavor profile

Every ingredient is represented by a vector of all of the PubChem id flavor compounds, as shown below in Table 3.6.

	Alias	**Molecules Vector**	**Category**
0	bakery products	[27457, 7976, 31252, 26808, 22201, 26331]	Bakery
1	bread	[1031, 1032, 644104, 527, 8723, 31260, 15394...]	Bakery
...	...	...	...
1788	egg	[6274, 5311110, 644104, 9609, 18827, 527...]	Animal Product
1789	olive oil	[6184, 31260, 5281168, 8103]	Additive

Table 3.6: Ingredients and their molecular vector and category

3.5.5 One Hot Encoding

Many machine learning algorithms only work with numerical data. One of the most common techniques to convert categorical data into numbers is the One Hot Encoding technique, which we will use for our second approach.

Another technique used in Integer Encoding is where each label is assigned a unique integer based on alphabetical ordering. For example, we have three categories: red, green, and blue.

We can change “red” to 1, “green” to 2, and “blue” to 3; this is called label encoding, which makes the encoding easily reversible.

The problem with this representation is that the model will probably capture a relationship between the three colors as blue ¿ green ¿ red which is not the case.

Instead, for this scenario, the one hot encoding is more suitable. It creates additional features based on the number of unique values in the categorical feature. Each color will be represented as a binary vector. We have three “colors,” so three binary variables are needed. A “1” value is placed in the binary variable for the color and a “0” value for the other colors.

$$\begin{aligned} Red &= [1, 0, 0] \\ Green &= [0, 1, 0] \\ Blue &= [0, 0, 1] \end{aligned}$$

Now, To train our model using the intersection of molecules, we will be applying a One Hot Encoding for every ingredient pair, but as we can see from the table above, the PubChem ID is of a bigger number. For example, the egg has a molecule with PubChem ID 5311110. Encoding this in binary would lead to high memory consumption, which is very computational. To get around that, we will use the index instead of using the PubChem ID as the identifier for the molecules. We know that there are only 1791 molecules in the dataset, so the one hot encoding vector for each ingredient molecule will have a length of 1791. To explain more, let us take salt as an example:

$$salt = [644104, 1130, 6202, 8094]$$

We will replace each PubChem ID with their index in the dataframe.

$$salt = [1397, 80, 131, 344]$$

Now encoding *salt* with binary value would be easy to do, if the index of an ingredient exists then its value is "1" otherwise it's "0".

$$salt = [0, ..1, \ldots, 1, .., 0]$$

3.5.6 Model (DNN)

For our second approach training model, the data inputs are the one hot encodings of the flavor compounds that two pairs of ingredients have in common—in addition, using the correlation value of the pair as their label as shown in Table 3.7.

Ingredient Pair Flavor compounds	**0**	**1**	**2**	...	**1789**	**1790**	**Correlation**
0	0	1	0	...	0	0	0.169771
1	0	0	1	...	0	0	0.124847
2	0	1	0	...	0	0	0.302941
...	...	...	...	...	...	...	...

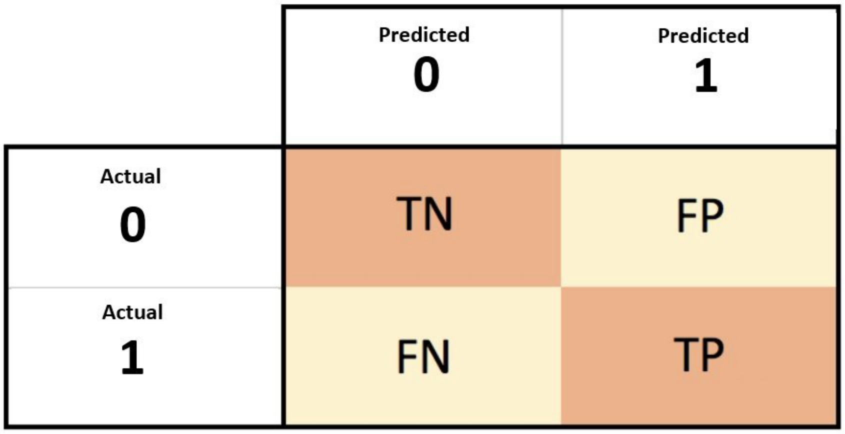
Table 3.7: One hot encodings flavor compounds of ingredient pairs

Each row represents flavor compounds in common with two pairs of ingredients, and each column is the index of one compound from the FlavorDB dataset. We have a total of 1791 flavor compounds. Moreover lastly, we used the exact correlation extracted from the binary matrix of recipes of all possible pairs of ingredients.

3.6 Evaluation's metrics

The essential thing in a machine learning problem is that our model performs well, and we can evaluate its performance by using the classification metrics below.

Confusion Matrix, also known as the error matrix as shown in figure 3.4, is a performance measure for classification tasks where we predict one or more classes. Its output is a table of the combination of the predicted and actual values. It consists of four blocks. Let us explain each block by taking our prediction model of ingredient pairs as our example:



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Figure 3.4: Confusion Matrix Table [44]

- **True Positive (TP):** TP represents the number of ingredient pairs that have been correctly classified as a good match.

- **True Negative (TN):** TN represents the number of ingredient pairs correctly classified as a bad match.

- **False Positive (FP):** FP represents the number of misclassified ingredient pairs that are a good match but are actually a bad match.

- **False Negative (FN):** FN represents the number of misclassified ingredient pairs that are a bad match but are actually a good match.

Precision measures the proportion of positively predicted labels that are actually correct. Mathematically, it represents the ratio of true positive to the sum of true positive and false positive[36].

$$Precision = TP/(TP + FP)$$

Recall represents the model's ability to correctly predict the positives out of actual positives. Mathematically, it represents the ratio of true positive to the sum of true positive and false negative[36].

$$RecallScore = TP/(FN + TP)$$

Accuracy is the ratio of true positives and true negatives to all positive and negative observations. Accuracy tells us how often we can expect our model will correctly predict the output out of the total number of times it made predictions.[36]

$$AccuracyScore = (TP + TN)/(TP + FN + TN + FP)$$

F1 Score represents the model score as a function of precision and recall score. It gives equal weight to both Precision and Recall for measuring its performance in terms of accuracy[36].

$$\begin{aligned} F1Score = 2 * PrecisionScore * RecallScore/ \\ (PrecisionScore + RecallScore) \end{aligned}$$

MAE is the mean of the absolute values of the individual prediction errors over all instances in the test set[51].

These are the metrics that we will use to evaluate the performance of our deep learning models.

3.7 Conclusion

In this chapter, we went through the global overview of the proposed approaches, explained the techniques used in each, and ended by explaining how we will evaluate our model.

Chapter 4

Test and Validation

4.1 Introduction

The aim of this work is building a deep learning model that will recommend food pairing and suggest combinations of ingredients to create innovative dishes. Two approaches are proposed: The one takes ingredient pairs frequency appearance in consideration, the second one is based on shared flavor compounds profile of the ingredient pairs.

To achieve that, we went through several stages that will explore in this chapter, alongside the datasets used for the training. We will present the set of the experiments conducted to validate the proposed approaches and the obtained results. In addition, we compare the proposed approaches by going through test cases with real scenarios.

4.2 Implementation Setup

In this section, we will present the tools and libraries we worked with and the development environment specifications .

4.2.1 Work Environment

In the development environment, we used Google Colaboratory [2], browser-based Python IDE that allows anyone to write and execute arbitrary Python code, which is especially useful for machine learning and data analysis. It offers free computing resources with GPU training acceleration.

4.2.2 Libraries

Python as our programming Language, a high-level language that contains many libraries that facilitated our data processing and building of our deep learning models.

We used the following Libraries:

- **Pandas** a software library written for the Python programming language for data manipulation and analysis.[7]
- **Numpy** is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.[6]
- **Matplotlib** a software library written for the Python programming language for data manipulation and analysis.[5]
- **Seaborn** a software library written for the Python programming language for data manipulation and analysis.[9]
- **Gensim** a software library written for the Python programming language for data manipulation and analysis.[1]
- **TensorFlow** a software library written for the Python programming language for data manipulation and analysis.[26]
- **Keras** is a high-level neural network library that runs on top of TensorFlow. It is more user-friendly because it's built-in Python.[4]
- **Scikit-learn** is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms.[8]

4.3 Dataset

In order to validate our work, we used three datasets: Simplified-Recipes-1M Dataset, Molecules dataset FlavorDB, and Traditional Algerian Dataset.

- **Simplified-recipes-1M Dataset:** A recipe-ingredient dataset contains 1,067,557 preprocessed recipes [52]. Figure 4.1 shows the statistics of the 1M+ recipes dataset and the distribution of the recipes over the different categories.

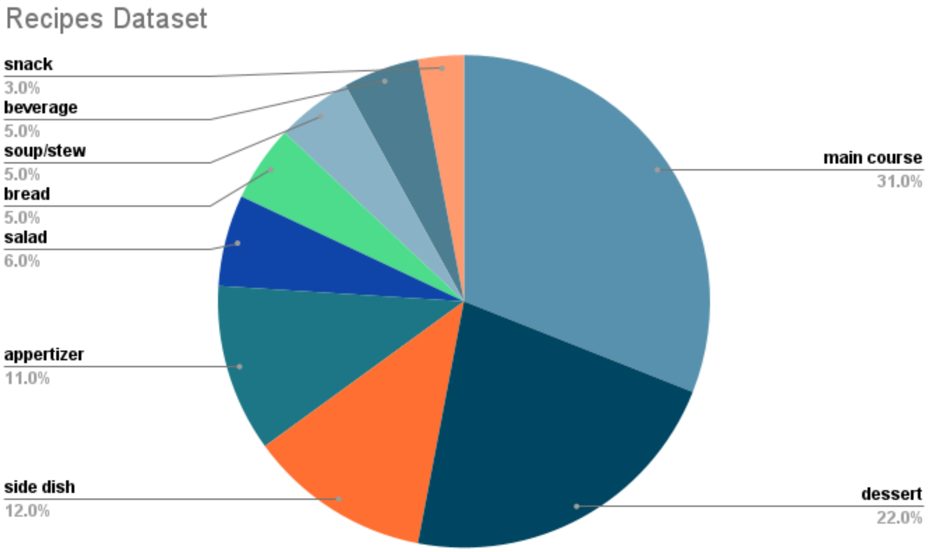


Figure 4.1: statistics of the 1M+ recipes dataset

- **Molecules Dataset:** FlavorDB is the dataset used to associate each ingredient with its respective flavor compounds FlavorDB comprises 25,595 flavor molecules representing an array of tastes and odors. Among these, 2254 molecules are associated with 940 natural ingredients belonging to 34 categories [45].

- **Traditional Algerian Dataset:** Recipes are collected from Ms. Bouayed Fatima Zohra's well-known Algerian book "la cuisine algérienne" [27] 323 recipes that consist of 232 main dishes recipes and 91 dessert recipes with a total of 116 unique ingredients from different regions in the country with different food categories such as desserts, bread, pastry, bread, etc. For the Algerian dataset, we will implement only the second approach because the Algerian dataset does not have enough recipes to train the SNN model properly. Thus we will show the results of the second approach only.

4.4 Data Processing

4.4.1 Data Processing: 1M+ Recipes

For the data processing, we worked with the 1M recipes dataset that was already processed and cleaned from raw messy recipes that looks like:

- 1 fennel bulb (sometimes called anise), stalks discarded
- 1/2-inch dice, and feathery leaves reserved for garnish
- 1 onion, diced
- 2 tablespoons unsalted butter
- 2 medium russet (baking) potatoes
- 2 cups chicken broth
- 1 1/2 cups milk

Processed to a list of string ingredients that looks like this :

[baking potatoes, butter, chicken, chicken broth, fennel, fennel bulb, garnish, leaves, milk onion, potatoes, unsalted butter]

Which made our task more manageable, But there were still many ingredient errors in the dataset where some ingredients are wrongly split, like in the example above; the ingredient fennel bulb is wrongly split into two ingredients and fennel bulb. And some are not considered natural ingredients, such as *extracted, dry, prepared, etc.*

Even after fixing the errors in the dataset, the issue is that for our second approach, the use of flavor compounds, we only have access to what FlavorDB has to offer. So for this approach to work, we can only use the 940 ingredients in FlavorDB. Some recipes are not usable since most of their ingredients, the molecular ingredient's profile does not exist in FlavorDB, so we got rid of them as well.

The lack of computational power also made using all of the 1M Recipes very resource-consuming. Therefore, we settled on cleaning and processing only 500,000 recipes; from these recipes, only 24,051 were valid. Table 4.1 below shows all of the statistics of our dataset, and Figure 4.2 below illustrates the stages of processing our data.

# of Correct Recipes	**Total # of Possible (known) pairs**	**Total # of unknown pairs**	**Total # Ingredients**
24,051 recipes	21,930 pairs	103,386 pairs	354 ingredients

Table 4.1: Statistics of known and unknown ingredient pairs and a total of unique ingredients

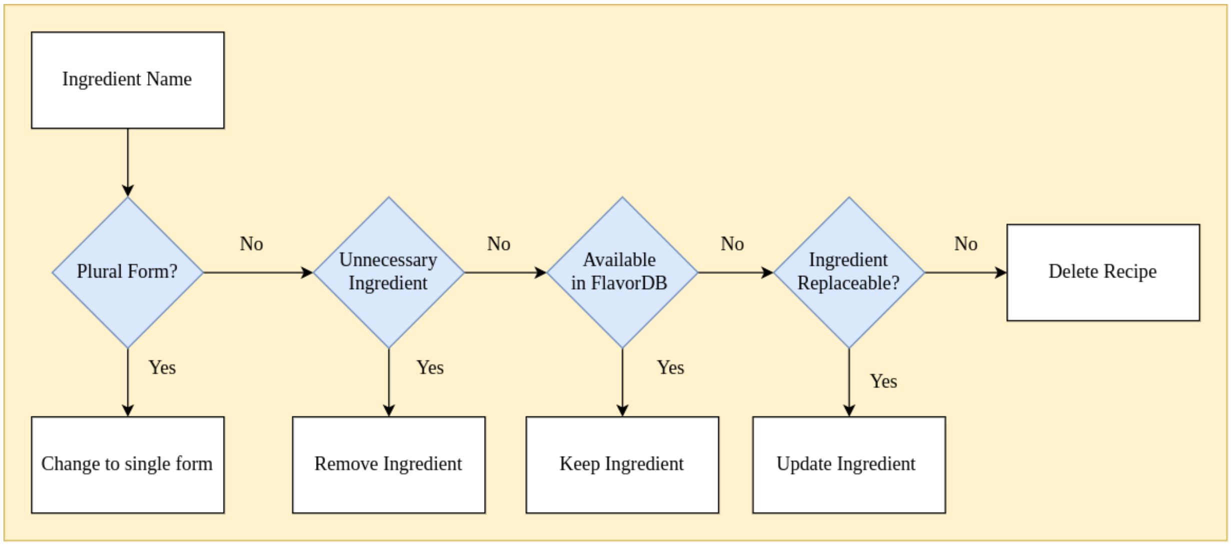


Figure 4.2: Stages of Dataset Cleaning and Processing Recipes

We then split Ingredient pairs: those with negative correlation are considered *unknown pairs*, and those with positive correlation as *known pairs*. Then, they are used to train our deep learning models. Table 2 shows the statistics of known and unknown ingredients and the number of unique ingredients.

4.4.2 Data Processing: Traditional Algerianne Dataset

The dataset contains 323 recipes with 116 ingredients. Already processed and cleaned by Kerbadj Tarek and Racha Chahboub in their master's thesis [34].

The dataset was collected from a Traditional Algerian Cook Book [27] where every recipe was in its raw format. Thus, the cleaning process consisted of many steps: removing all the unnecessary information and keeping only the existing unique ingredient in each recipe. Furthermore, all ingredients used needed to have their flavor compounds in FlavorDB. The missing ingredients needed to be replaced with other relevant ingredients, and the ones that could not be replaced were deleted as well as their recipes. The resulting final dataset consists of 323 recipes with 116 unique ingredients.

To label the dataset, we will also split ingredient pairs into *known and unknown pairings*. Then use the known ingredient pairs to train our deep learning model following the second approach only because the data is too small to allow the training of the SNN architecture proposed in the first approach.

4.5 Evaluation of the first Approach

4.5.1 Feature Extraction: Word Embedding

For training our model, we represented every unique ingredient with a **50-dimensional vector of real numbers**. Figure 4.3 below shows a 3D visualization of an example of some similar ingredients after using a word embedding on our ingredients.

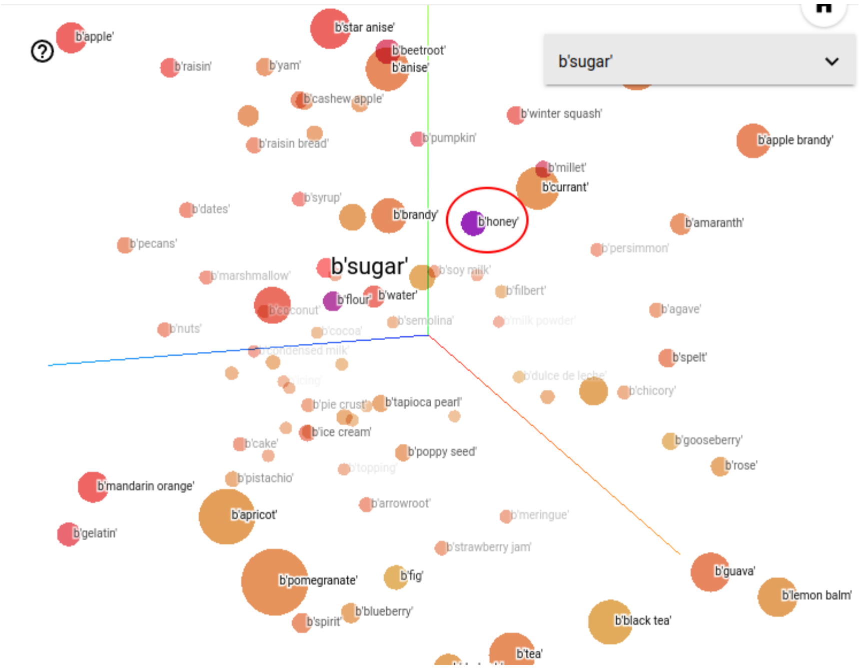
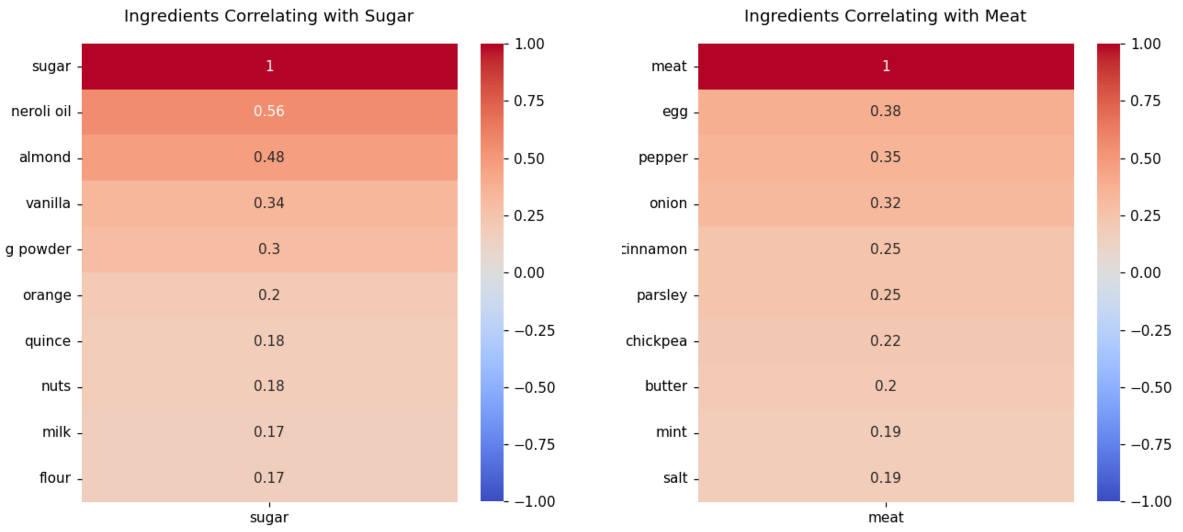


Figure 4.3: 3D Visualization of all the ingredients similar/close to Sugar using word embedding (1M+ recipes)

We can see that the closer ingredients in the 3D graph to sugar, the more similar they are. For example, *honey and sugar, sugar and mandarin orange.* We can say that these three pairs are pretty similar since they are considered sugary ingredients.

4.5.2 Correlation

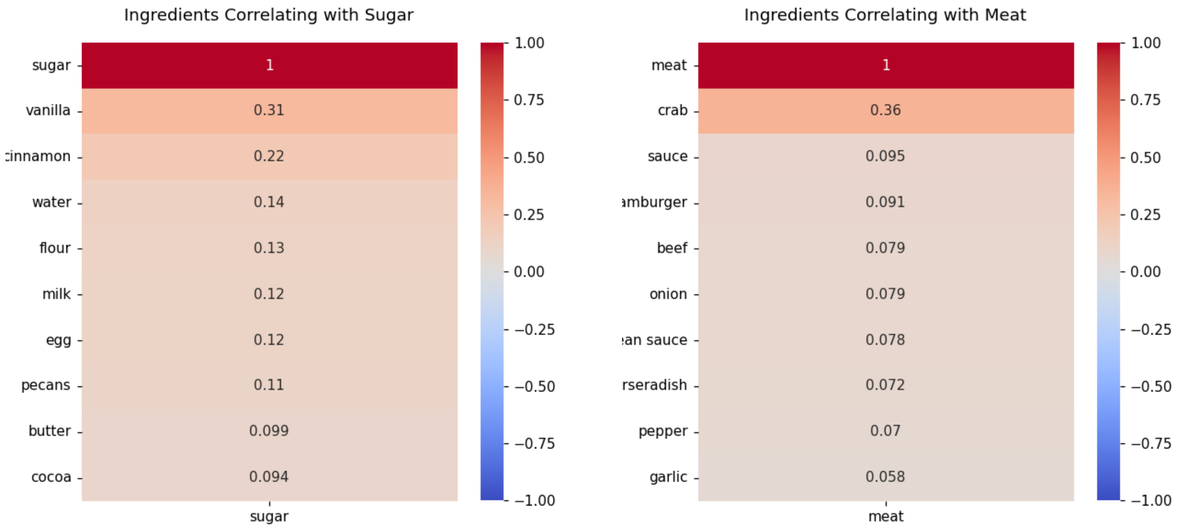
We calculate the correlation between ingredients. Figures 4.4 and 4.5 below shows heat maps of the correlation results for both **1M+ Recipes Dataset** and the **Traditional Algerian Dataset**.



(a) Heatmap that shows the top 10 ingredients correlated with Sugar in the Algerian Dataset

(b) Heatmap that shows the top 10 ingredients correlated with Meat in the Algerian Dataset

Figure 4.4: Correlation Heatmaps of the Algerian Dataset



(a) Heatmap that shows the top 10 ingredients correlated with Sugar in the 1M+ recipes

(b) Heatmap that shows the top 10 ingredients correlated with Meat in the 1M+ recipes

Figure 4.5: Correlation Heatmaps of the 1M+ Recipes Dataset

The correlation visualizations of sugar and meat presented in Figures 4.4a and 4.4b above show the nature of the relationship with other ingredients. The higher the correlation between two ingredients, the better the match of the ingredient pair. For example, it shows *sugar and almond, sugar and vanilla, sugar and baking powder, meat and egg, meat and pepper, meat and onion.* All of these pairs are frequently used in today's cuisine.

We also note that the correlated ingredients with sugar are very different in each dataset since the 1M+ recipes are more worldwide data; in contrast, the Algerian results are specific to the traditional country's cuisine style. We can illustrate the difference more in Figures 4.6 and 4.7, which show an enormous difference in the pairs of the ingredients with the highest correlation, for example, in the **1M+ recipes**. We note *vanilla and milk, vanilla and sugar* are the most used ingredient pairs. In contrast, in the **Traditional Algerian dataset** we note *pepper and onion, pepper and cooking oil* are the most used ones. Most of the traditional algerian recipes have much spice in them, for example, *zviti, dobara, etc.* [49], that would explain why pepper is used so much since it is considered a spicy ingredient primarily used for seasoning [40].

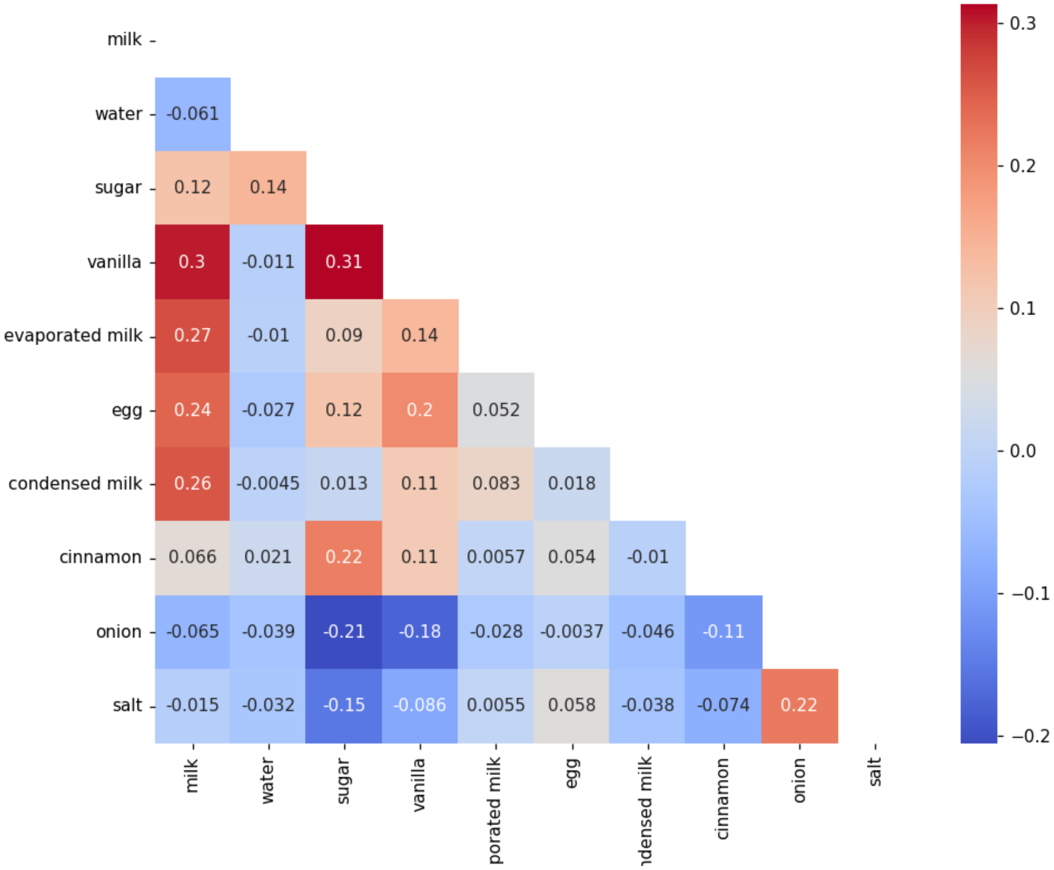


Figure 4.6: correlation heatmap to understand the linear relationship between two pairs of ingredients in the 1M+ recipes

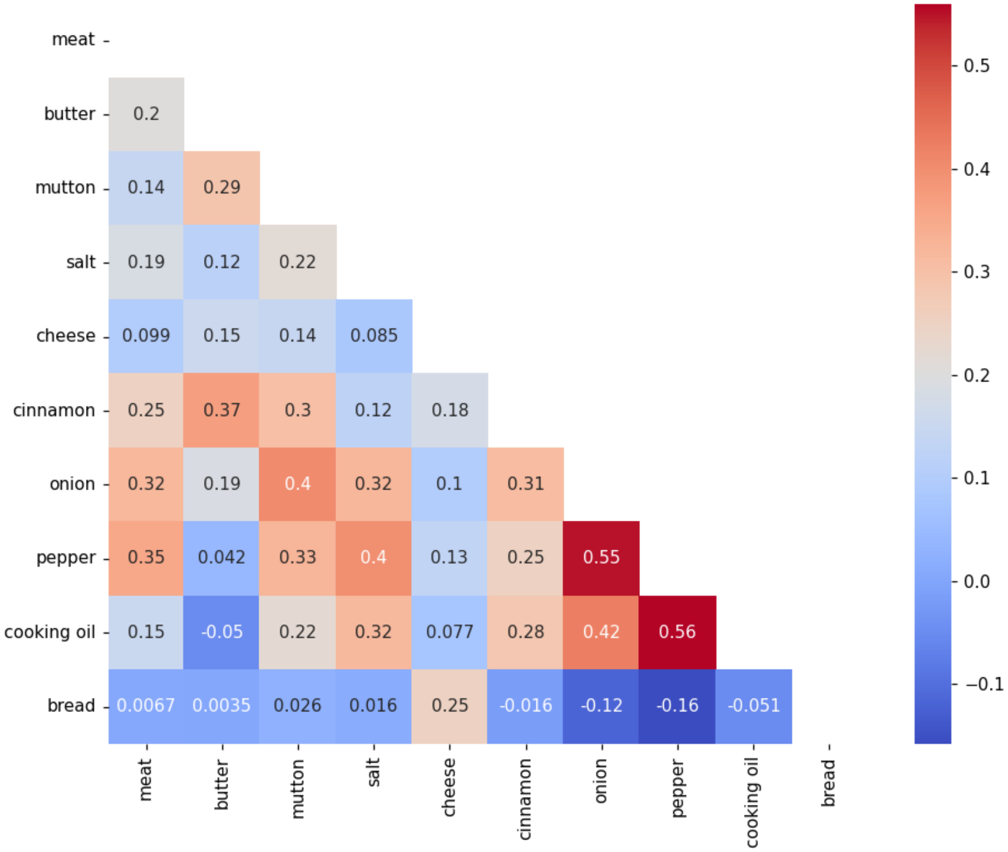
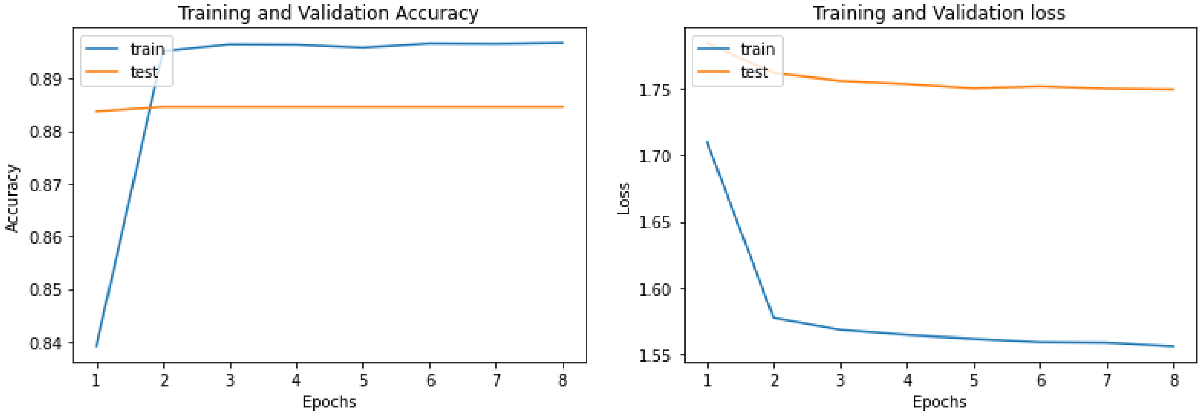


Figure 4.7: correlation heatmap to understand the linear relationship between two pairs of ingredients in the Algerien dataset recipes

4.5.3 Results: SNN First Approach Model Evaluation

After our model was trained with the 21,930 known pairs, we used a threshold for our correlation of 0.3 to have a classification task. If the correlation is higher than the threshold, the value is 1; otherwise, it is 0. Both approaches will have the same threshold to have a comparative analysis. We evaluate our model using the most common metrics that will let us confirm that our model was trained properly. We note the following results shown in Figure 4.10 and Table 4.4. The model gives a good results with a precision, recall and F1 scores of 0.98, 0.89, 0.93, respectively. In order to verify the efficacy of the results, we compare the similarity values generated by the SNN model with the correlation calculated on the dataset by using mean absolute error (MAE) that gives 0.11.



(a) Train/Validation Accuracy SNN Model (b) Train/Loss Accuracy SNN Model

Figure 4.8: SNN Training and Validation Accuracy and Loss Plots with the 1M+ Recipes Dataset

Metric	**Value**
Accuracy	0.8974
Precision	0.9847
Recall	0.8974
F1 score	0.9333
MAE	0.1172

Table 4.2: Results on the 1M+ recipes dataset using the SNN approach

4.6 Evaluation of the Second Approach

4.6.1 Feature Extraction: Ingredient's Molecular Profile

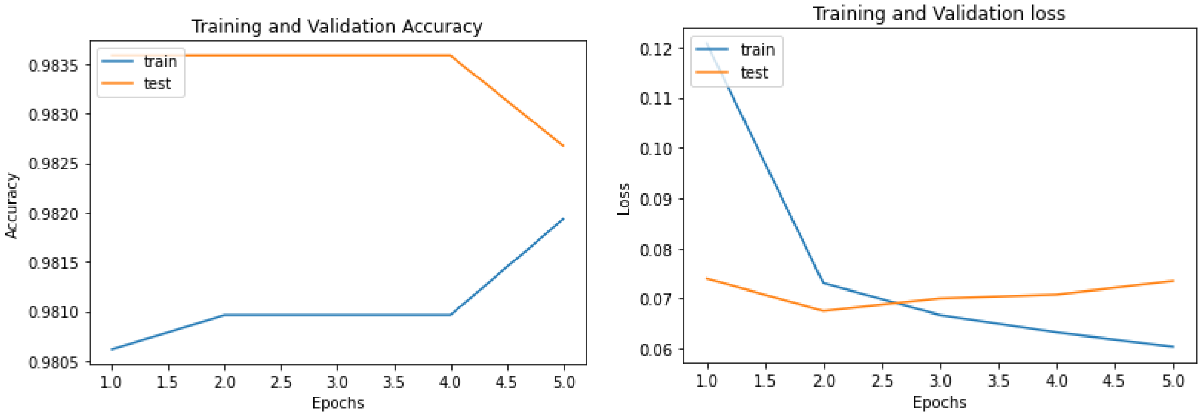
The **1M+ Recipe** dataset was already ready to use after we had cleaned it and processed it for our first approach. Each ingredient was represented by the set of molecules that constituted it. Thus, we will get the molecular profile for each ingredient. Then, we use the one-hot encoding to map the ingredient's molecular profile to a numeric array. The resulting array was a 1791 dimensional vector.

We apply the same process for the **Traditional Algerian Dataset** in extracting the ingredient pairs molecules and applying one hot encoding for each pair with a 1791 dimensional vector.

Now that we have our inputs ready for both datasets, we will train our model on both datasets (1M+ recipes and Algerian Dataset) and evaluate the results.

4.6.2 Results: DNN Second Approach Model Evaluation: 1M+ Recipes Dataset

We used a threshold of 0.3 for our correlation. We note the following results:



(a) Train/Validation Accuracy DNN Model

(b) Train/Loss Accuracy DNN Model

Figure 4.9: DNN Training and Validation Accuracy and Loss Plots with the 1M+ Recipes Dataset

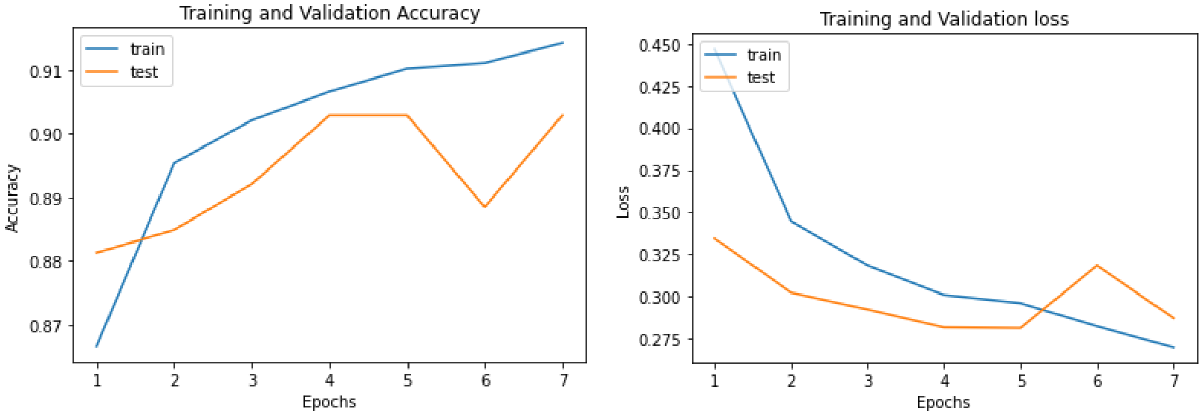
Metric	**Value**
Accuracy	0.9835
Precision	0.9765
Recall	0.9835
F1 score	0.9770
MAE	0.0336

Table 4.3: Results on the 1M+ recipes dataset using the DNN approach

To verify the efficacy of the results, we compare the the DNN model output without applying the threshold with the correlation calculated on the dataset by using mean absolute error (MAE) that gives 0.03. The model gives very good results with a precision, recall and F1 scores of 0.98, 0.97, 0.97, respectively.

4.6.3 Results: Second Approach DNN Model Evaluation: Algerian Dataset

We used the same threshold of 0.3 for our correlation. We note the following results:



(a) Train/Validation Accuracy DNN Model

(b) Train/Loss Accuracy DNN Model

Figure 4.10: DNN Training and Validation Accuracy and Loss Plots with the Algerian Dataset

Metric	**Value**
Accuracy	0.8855
Precision	0.9764
Recall	0.8906
F1 score	0.9291
MAE	0.1258

Table 4.4: Results on the Algerian Dataset using the DNN approach

4.7 Discussion

To evaluate our prediction results of ingredient pairs, we will test the relevancy of our ingredient pairs. Firstly, let us verify that if ingredient A matches with ingredient B, similar ingredients to A would most likely pair well with ingredient B. We can check existing recipes that use those pairs to confirm our proposition.

Secondly, we will compare the predicted ingredient pairs results of the frequency model and molecule profile approaches.

4.7.1 Testing on 1M+ Recipes

To demonstrate our models' accuracy and verify if the resulting predictions of food ingredient pairings are accurate, we will analyze unknown pairings returned by our system.

To test our first proposition, we randomly chose one ingredient pair: *skimmed milk and honey*; they are considered a known pair commonly used. Then we picked four more ingredients similar to *skimmed milk* since they are all considered dairy products *(buttermilk,* *gruyere cheese, romano cheese, soy yogurt)* as shown in Table 4.5 to constitute a pair with *honey.* The prediction results of all four pairings were consistently high even though they were considered uncommon pairs before. To confirm if the pairs predicted are used, we researched if there are any recipes in worldwide cuisine that use those same ingredient pairs, and it was the case as shown in Table 4.6.

Similar Ingredients	**Unknown Pairing**	**Predicted Value (SNN)**
Buttermilk	-0.005	1.0
Gruyere cheese	-0.01	1.0
Romano cheese	-0.007	1.0
Soy yogurt	-0.002	1.0

Table 4.5: Examples of unknown pairings and their predicted scores

Ingredient Pairs	**Recipes**
Buttermilk & Honey	Buttermilk Honey Bread [11]
Gruyere cheese & Honey	Gruyere and Honey Sandwich [46]
Romano (pecorino) cheese & Honey	Fried pecorino with honey and summer herbs [63]
Soy yogurt & Honey	Soy yogurt with honey [13]

Table 4.6: Examples of recipes that use the unknown pairings predicted

For our comparative analysis between the first approach prediction results using the frequency of ingredient pairs appearance and the second approach using the shared flavor compounds between the ingredients of the pair, we picked four widely used ingredients *(rice, cinnamon, milk, vanilla).* Then, we retrieved the top 6 ingredient pairings predicted on both SNN Model and the DNN Model. Based on observations, our SNN prediction model recommended ingredients used frequently in everyday cooking, for example, *(rice and salad dressing, rice and swiss cheese, cinnamon and pudding, cinnamon and potato).* On the other hand, the DNN recommended ingredient pairs with many common chemical compounds. However, some of those recommendations are not considered good ingredient pairs as they are rarely used together in recipes, for example, *(rice and coffee, cinnamon and cherry pepper).* The results are shown below in Table 4.7 and Table 4.8.

	rice		**cinnamon**	
Rank	**1st Approach (SNN)**	**2nd Approach (DNN)**	**1st Approach (SNN)**	**2nd Approach (DNN)**
1	salad dressing	soybean	pudding	cherry pepper
2	lettuce	buckwheat	cheddar cheese	basil
3	relish	coffee	potato	pepper
4	ginger	filbert	rosemary	jasmine
5	mint	caviar	cucumber	dumpling
6	swiss cheese	sweetcorn	pomegranate	bonito

Table 4.7: a comparative analysis between first and second approach ingredient pairs results (1)

	milk		**vanilla**	
Rank	**1st Approach (SNN)**	**2nd Approach (DNN)**	**1st Approach (SNN)**	**2nd Approach (DNN)**
1	shrimp	beer	shrimp	tea
2	leek	caviar	shallot	olive
3	shallot	bonito	eggplant	corn oil
4	chicken	tea	leek	white wine
5	broccoli	yogurt	sesame	Cherry pepper
6	sesame	cider	chicken	tomato

Table 4.8: A comparative analysis between first and second approach ingredient pairs results (2)

From these results, we can argue that humans experiencing food makes it possible to know what ingredients combination works and does not. The SNN model proposes combinations of ingredients based on their frequency of appearance in a particular cuisine; this is used for recommending a familiar dish to customers based on their preferences. However, it makes discovering new combinations of food difficult as the new combinations are not already known. On the other hand, the DNN model helps the chefs to innovate because it proposes a list of new pairs to experiment with and validate with customers. Rather than establishing new pairs manually in an ample complex space of ingredients combination, the DNN model, based on the food pairing hypothesis, returns a list of possible success combinations that share a similar molecular profile.

4.7.2 Testing on Algerian Recipes Dataset

To validate the conclusion deducted above, we took two ingredients, *tomato, and cheese,* and analyzed the pairs proposed by our system matching those latter. For that purpose, we keep their top 5 ingredient pairs predicted as shown in Table 4.9 below.

Ingredient	Top 5 Ingredient Pairs
Tomato	melon
	banana
	apple
	grape
	cherry
Chesse	cottage cheese
	gruyere cheese
	tangerine
	apple
	nuts

Table 4.9: The proposed pairs for tomato and cheese ingredients by DNN model

We can observe that the resulting ingredient pairs share a large flavor compound in common, but in reality, some of those pairs would seem peculiar for some cuisine styles. For example, apple and cheese; one is sweet, the other is savory, and some would say that would not work. However, it is quite the opposite since apple and cheese are considered the perfect pair and have an irresistible taste as they offer the perfect marriage of sweet and savory [64].

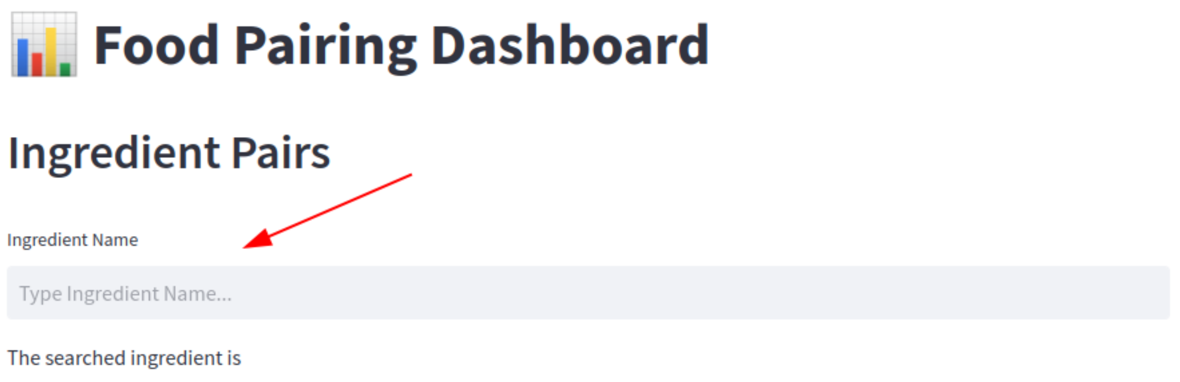
Additionally, there was a long debate on 'tomato.' Some consider it a vegetable, and some consider it a fruit. The latter consideration is related to the science of botany, where the fruit is a ripened flower ovary with seeds. That is why tomatoes are classified as fruit because they contain seeds and grow from the flower of the tomato plant [33] [48]. We can see why it is a fruit since tomatoes pair well with fruits *(melon, banana, apple, grape, cherry)*.

These examples validate that our model can help chefs discover new innovative ingredient pairs to create new recipes, reducing the search for new ingredient combinations that would please food lovers worldwide.

4.8 Application Interface

We created a dashboard to use our food pairing solution for all type of users.

First the user will be prompt to the search engine of our food pairing platform as shown in Figure 4.11.



Food Pairing Dashboard

Ingredient Pairs

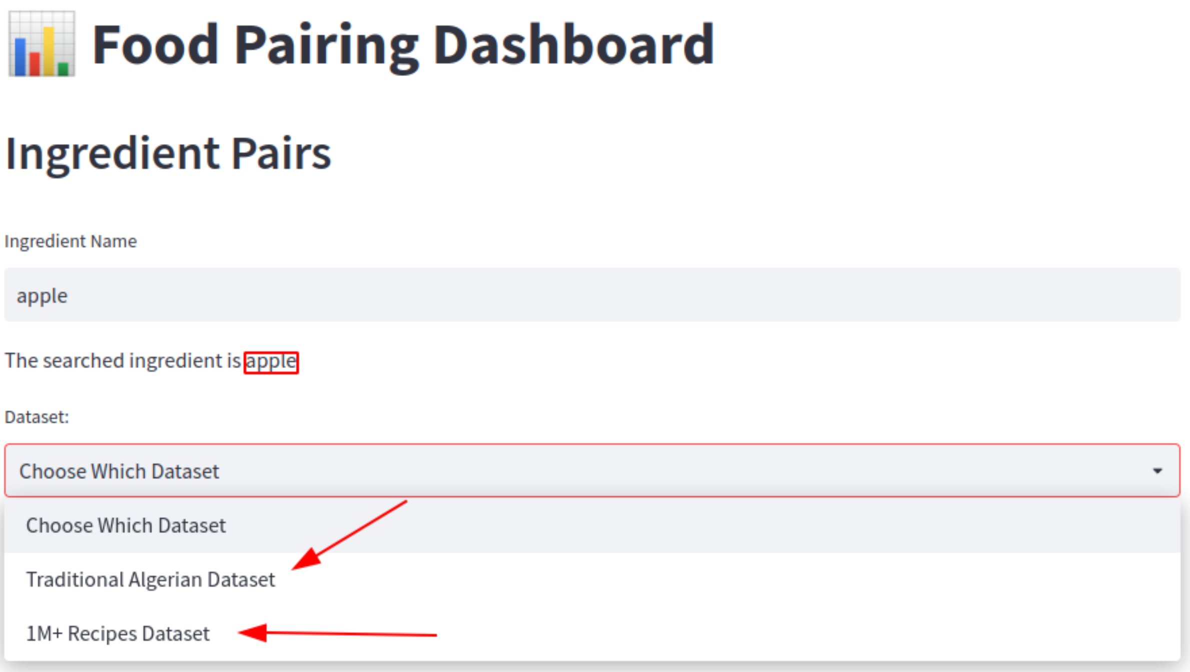
Ingredient Name

Type Ingredient Name...

The searched ingredient is

Figure 4.11: Food Pairing Search Engine

Next, the user has the ability to choose either one of the datasets *(1M+ Recipes Dataset, Traditional Algerian Dataset)* as shown in Figure 4.12.



Food Pairing Dashboard

Ingredient Pairs

Ingredient Name

apple

The searched ingredient is apple

Dataset:

Choose Which Dataset

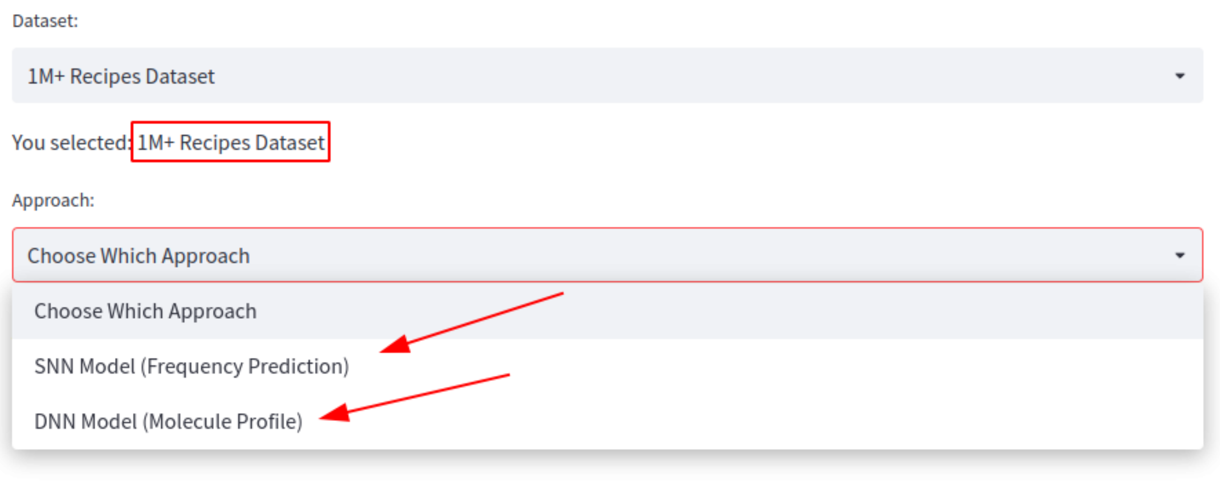
Choose Which Dataset

Traditional Algerian Dataset

1M+ Recipes Dataset

Figure 4.12: Choosing Dataset Select Box

For the 1M+ Recipes Dataset, the user can choose *the SNN Model (Frequency Prediction) Approach or the DNN Model (Molecule Profile) Approach*, as shown in Figure 4.13.



Dataset:

1M+ Recipes Dataset

You selected: 1M+ Recipes Dataset

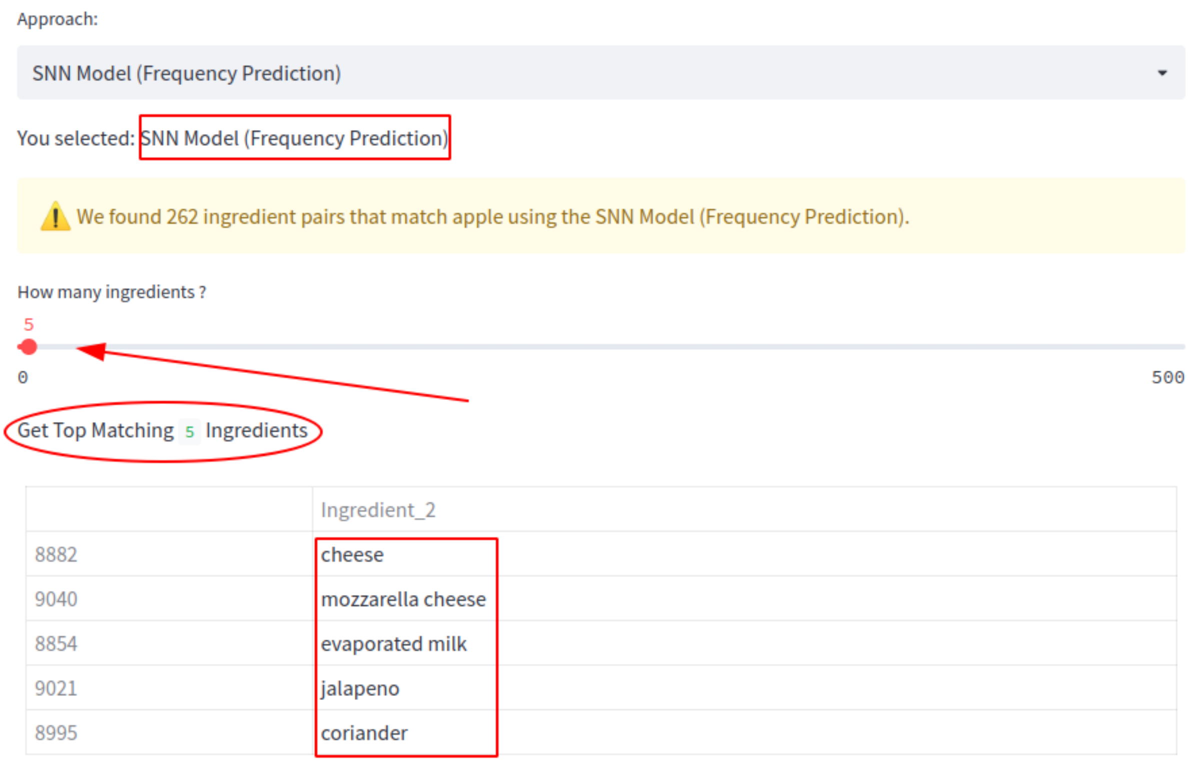
Approach:

Choose Which Approach

Choose Which Approach
SNN Model (Frequency Prediction)
DNN Model (Molecule Profile)

Figure 4.13: Choosing Approach Select Box

After choosing the dataset type and the approach type, the results show the Top Matching Ingredients, where there is a slider that lets the user control the number of ingredient pairs printed, as shown in Figure 4.14 and Figure 4.15.



Approach:

SNN Model (Frequency Prediction)

You selected: SNN Model (Frequency Prediction)

⚠ We found 262 ingredient pairs that match apple using the SNN Model (Frequency Prediction).

How many ingredients ?

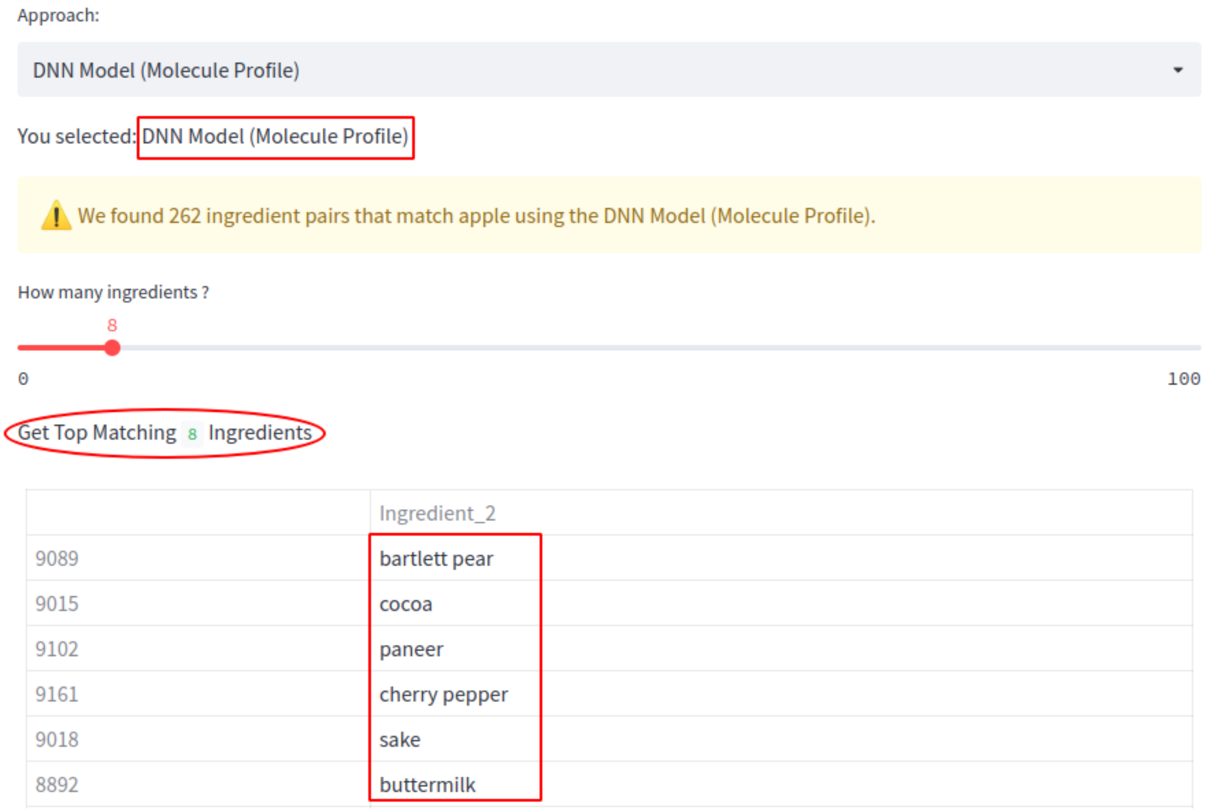
5

0 500

Get Top Matching 5 Ingredients

	Ingredient_2
8882	cheese
9040	mozzarella cheese
8854	evaporated milk
9021	jalapeno
8995	coriander

Figure 4.14: Results of the SNN Model (1M+ Recipes)



Approach:

DNN Model (Molecule Profile)

You selected: DNN Model (Molecule Profile)

⚠ We found 262 ingredient pairs that match apple using the DNN Model (Molecule Profile).

How many ingredients ?

8

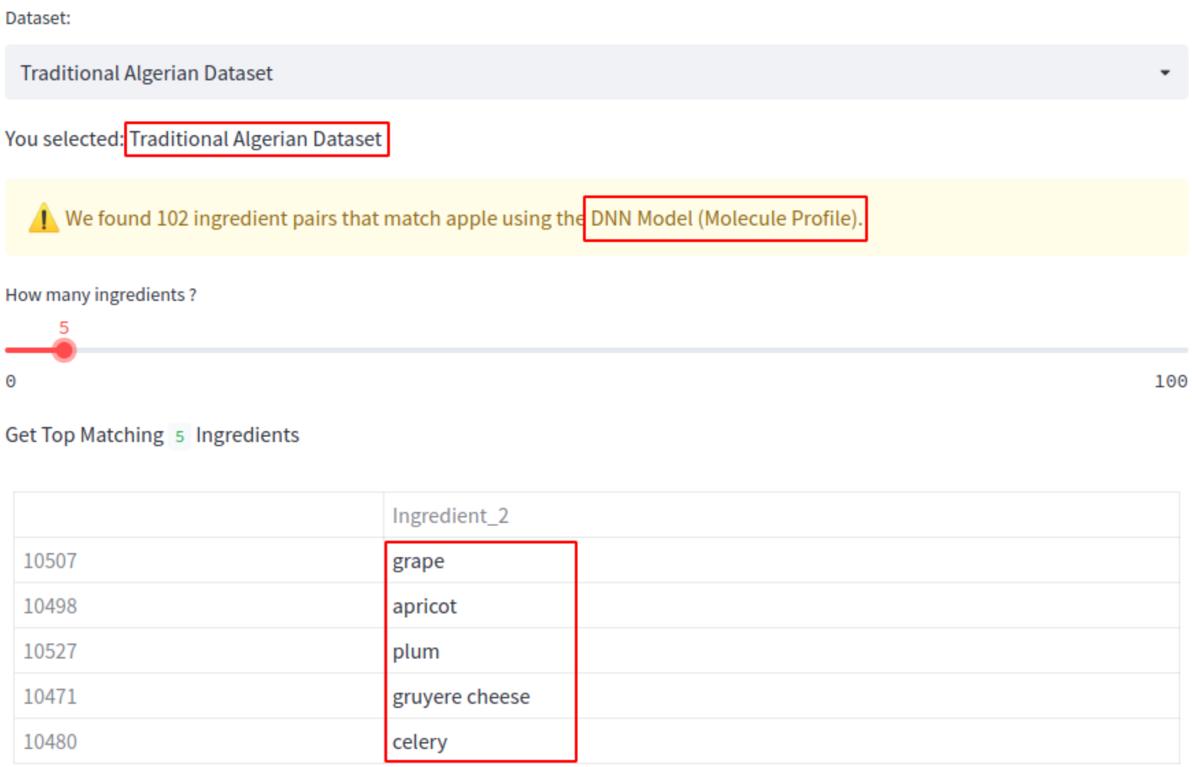
0 100

Get Top Matching 8 Ingredients

	Ingredient_2
9089	bartlett pear
9015	cocoa
9102	paneer
9161	cherry pepper
9018	sake
8892	buttermilk

Figure 4.15: Results of the DNN Model (1M+ Recipes)

If the user chooses the *Traditional Algerian Dataset*, only the *DNN Model (Molecule Profile)* is available, as shown in 4.16.



Dataset:

Traditional Algerian Dataset

You selected: Traditional Algerian Dataset

⚠ We found 102 ingredient pairs that match apple using the DNN Model (Molecule Profile).

How many ingredients ?

5

0 100

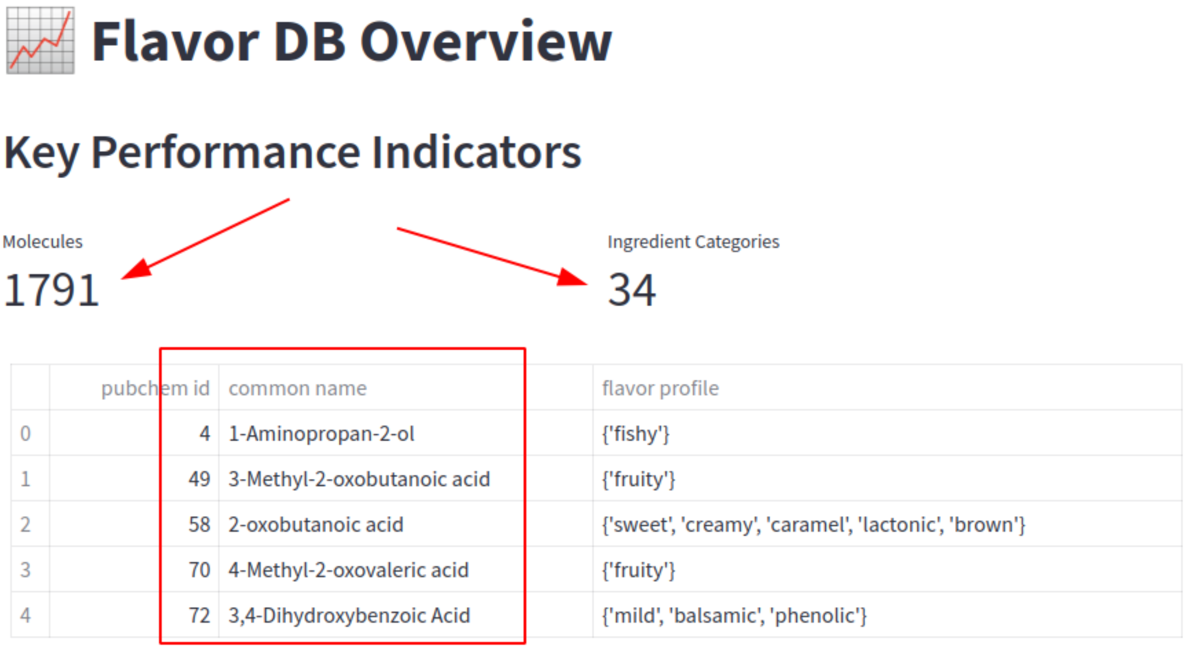
Get Top Matching 5 Ingredients

	Ingredient_2
10507	grape
10498	apricot
10527	plum
10471	gruyere cheese
10480	celery

Figure 4.16: Results of the DNN Model (Traditional Algerian Recipes)

Figure 4.17 and 4.18 illustrate an overview on the FlavorDB Datasets. It shows the Key Performance Indicators that consist of 1791 molecules belonging to 34 Categories.

Each molecule is identified by a *PubChem ID* with its flavor profile as shown in 4.17, and each Ingredient has its molecular vector profile, which consists of all the *PubChem ID* instead of the common name as shown in 4.18.



📈 Flavor DB Overview

Key Performance Indicators

Molecules

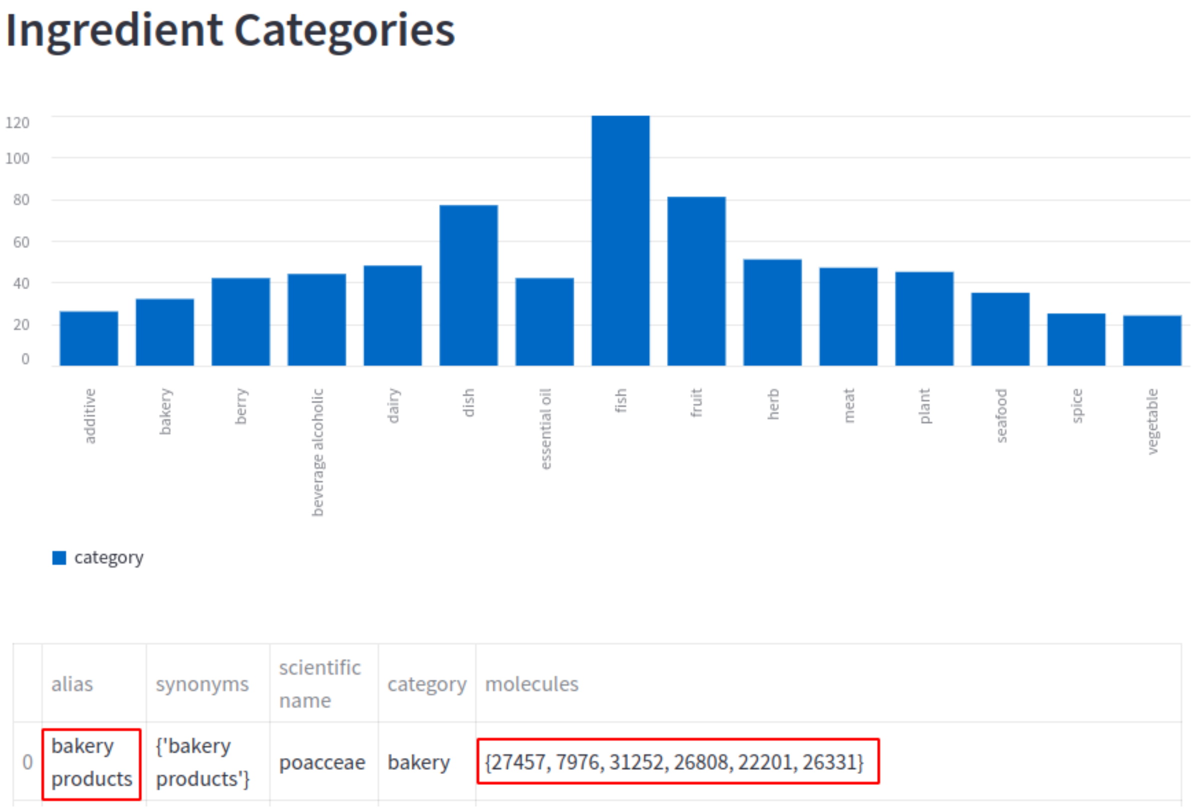
1791

Ingredient Categories

34

	pubchem id	common name	flavor profile
0	4	1-Aminopropan-2-ol	{'fishy'}
1	49	3-Methyl-2-oxobutanoic acid	{'fruity'}
2	58	2-oxobutanoic acid	{'sweet', 'creamy', 'caramel', 'lactonic', 'brown'}
3	70	4-Methyl-2-oxovaleric acid	{'fruity'}
4	72	3,4-Dihydroxybenzoic Acid	{'mild', 'balsamic', 'phenolic'}

Figure 4.17: Flavor DB Overview 1



Ingredient Categories

	alias	synonyms	scientific name	category	molecules
0	bakery products	{'bakery products'}	poacceae	bakery	{27457, 7976, 31252, 26808, 22201, 26331}

Figure 4.18: Flavor DB Overview 2

4.9 Conclusion

In this chapter, we tackled all the steps in detail that enabled us to train our deep learning models and went through the evaluation of our models to test their performance. In addition, to validate our work, we tested our models in actual data with several study cases that helped us discuss the results and how this research can help future food science studies.

General Conclusion

The food industry is one of the biggest industries worldwide. The increase in population will increase the supply and demand rate and increase jobs in the food sector. It will be an opportunity for chefs to rise above the status quo with their food pairing innovation to beat their competitors.

The purpose of our work is to study the field of food pairing. For this aim, we propose two different approaches. The first takes the frequency of appearance of ingredient pairs, and the second is based on the food pairing hypothesis of ingredient pairs that share similar molecular flavors in two different recipes. The 1M+ recipes give us a worldwide perceptive since it incorporates different cuisine styles from different regions of the globe. The second dataset is specific to the traditional algerian cuisine style recipes.

We arrived from the comparative analysis of both approaches that both are necessary to help us enrich the culinary world. The first approach will help us recommend ingredient pairs familiar to the customers' preferences, and it is left to the chefs to create innovative dishes from those recommendations. The second approach will recommend unique and new ingredient pairs that share similar molecular flavors without the need for chefs to keep trying different ingredients without a good reason. Thus chefs will gain time and resources directed solely to their creativity and imagination.

These conclusions were possible using the natural language processing field and its many techniques and deep learning models using neural networks trained on a cleaned and processed corpus of recipes. We then evaluated the models to check their performance using evaluation metrics that tell us if our model was appropriately trained. We tested our models on the unknown ingredient pairs with a negative correlation. Finally, we discussed the results provided to us by our prediction models, and to validate the results, we tested our models through several test cases.

Our future perspective in this work is to test our approaches with different datasets from different regions and coutries to have richer data that will give us more insights into the art of food pairing. We will put our work to the test by chefs that want to use our food pairing solution to create innovative recipes; thus, this will help us to customize our approach and make it more suitable on the business level.

References

[1] Gensim: Topic modelling for humans. Software available from radimrehurek.com, Last accessed July 2022.

[2] Google colaboratry platform. Software available from colab.research.google.com, Last accessed July 2022.

[3] How to spherify. From the Great British Chefs Blog, Last accessed September 2022.

[4] Keras: the python deep learning api. Software available from keras.io, Last accessed July 2022.

[5] Matplotlib: Visualization with python. Software available from matplotlib.org, Last accessed July 2022.

[6] Numpy. Software available from numpy.org, Last accessed July 2022.

[7] Pandas. Software available from pandas,pydata.org, Last accessed July 2022.

[8] Scikit-learn: Machine learning in python. Software available from scikit-learn.org, Last accessed July 2022.

[9] Seaborn: statistical data visualization. Software available from seaborn.pydata.org, Last accessed July 2022.

[10] Deepmind technologies, 2010. Last accessed August 2022.

[11] Buttermilk honey bread, 2021. Published in allrecipes Website on www.allrecipes.com, Last accessed September 2022.

[12] Correlation and regression, .n.d. Published in BMJ.

[13] Soy yogurt recipe, .n.d. Published in overcomingms Website on www.overcomingms.org, Last accessed September 2022.

[14] What is correlation?, .n.d. Published in Statistical Discovery LLC, Last accessed September 2022.

[15] Ahnert S. Bagrow J. et al. Ahn, YY. Flavor network and the principles of food pairing. *Sci Rep*, 1(196), 2011.

[16] Ahnert S. Bagrow J. et al. Ahn, YY. Exploring the food pairing hypothesis in saudi cuisine using genetic algorithm. *Mathematical Problems in Engineering*, vol.(Article ID 3627715, 16 pages), 2021.

[17] Pragati Baheti. Activation functions in neural networks [12 types use cases], 2022. Last accessed September 2022.

[18] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer, 2006.

[19] Heston Blumenthal. *The Big Fat Duck Cookbook.* Bloomsbury USA; Slp edition, 2008.

[20] Jason Brownlee. 14 different types of learning in machine learning, 2019. Last accessed August 2022.

[21] Andrew G. Barto by Richard S. Sutton. *Reinforcement Learning, second edition: An Introduction (Adaptive Computation and Machine Learning series).* Bradford Books; second edition, 2018.

[22] Remi Cadene. Master's thesis deep learning for visual recognition. 2016. Last accessed September 2022.

[23] Yonggyu Park Jungwoon Shin Jaewoo Kang Donghyeon Park, Keonwoo Kim. Kitchenette: Predicting and recommending food ingredient pairings using siamese neural networks. 2019.

[24] IBM Cloud Education. Natural language processing (nlp), 2020. Last accessed September 2022.

[25] IBM Cloud Education. Neural networks, 2020. Last accessed September 2022.

[26] Martin Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems.

[27] Bouayed Fatima-Zohra. *La cuisine algérienne.* Temps Actuels, January 1970.

[28] Stephanie Glen. Correlation coefficient, .n.d. From IIT JEE Study Material, Last accessed September 2022.

[29] Stephanie Glen. Correlation coefficient: Simple definition, formula, easy steps, .n.d. From StatisticsHowTo.com, Last accessed September 2022.

[30] Hannah Herrera. What is molecular gastronomy?, n.d. Last accessed August 2022.

[31] Aaron Courville Ian Goodfellow, Yoshua Bengio. *Deep Learning (Adaptive Computation and Machine Learning series).* The MIT Press; Illustrated edition, 2016.

[32] N. Rochester C.E. Shannon J. McCarthy, M. L. Minsky. A proposal for the dartmouth summer research project on artificial intelligence, 1955. Last accessed September 2022.

[33] Rachel A. Fisher Jean A.T. Pennington. Classification of fruits and vegetables. *Journal of Food Composition and Analysis 22S*, 2019. Last accessed September 2022.

[34] Bacha S. Karbedj T., Chaboub R. Evolutionary algorithm for the study of the food pairing hypothesis in the algerian cuisine. *Master's Thesis, University of Blida 1*, June 2022.

[35] Hosokawa C. Matsushima K. Varshney LR. Kazama M., Sugimoto M. and Ishikawa Y. A neural network system for transformation of regional cuisine style. 2018.

[36] Ajitesh Kumar. Accuracy, precision, recall f1-score – python examples, 2022. Posted in Data Science, Machine Learning, Python., Last accessed September 2022.

[37] Niklas Lang. Stemming vs. lemmatization in nlp, 2022. Published in Towards Data Science, Last accessed September 2022.

[38] Ben Lutkevich. natural language processing (nlp), 2021. Last accessed September 2022.

[39] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images, 2019. Last accessed July 2022.

[40] MasterClass. Cooking with black pepper: Understanding how black pepper modifies the flavor of food, 2021. Published in MasterClass Website on www.masterclass.com, Last accessed September 2022.

[41] MasterClass. A guide to molecular gastronomy: 8 molecular gastronomy methods, 2021. Last accessed August 2022.

[42] MasterClass. What is emulsification and how does it work? plus how to fix broken emulsions, 2021. Last accessed August 2022.

[43] Sanatan Mishra. Unsupervised learning and data clustering, 2017.

[44] Bilal Mussa. A python function to get all the possible stats from a confusion matrix, 2022. Published in Towards Dev, Last accessed September 2022.

[45] Rudraksh Tuwani Rakhi NK Shubham Dokania Arvind Iyer Ayushi Gupta Shubhra Agrawal Navjot Singh Shubham Shukla Kriti Kathuria Rahul Badhwar Rakesh Kanji Anupam Jain Avneet Kaur Rashmi Nagpal Ganesh Bagler Neelansh Garg, Apuroop Sethupathy. Flavordb: a database of flavor molecules. 46:D1210–D1216, 2018. Last accessed September 2022.

[46] Nick. Gruyere and honey sandwich, 2009. Published in macheesmo Website on www.macheesmo.com, Last accessed September 2022.

[47] Johan Langenbick Peter Coucquyt, Bernard Lahousse. *The Art and Science of Foodpairing: 10,000 flavour matches that will transform the way you eat.* Firefly Books; 1st edition, 2020.

[48] Melissa Petruzzello. Is a tomato a fruit or a vegetable?, .n.d. Published in Britannica Website on www.britannica.com, Last accessed September 2022.

[49] Rochdi Rais. Top 25 most popular foods in algeria – top algerian dishes, 2021. Published in Chef's Pencil's Website on www.chefspencil.com, Last accessed September 2022.

[50] Alan L.Kelly Róisín Burke, Hervé Thism. Molecular gastronomy. 2016.

[51] Webb G.I. Sammut, C. Mean absolute error, 2011. Last accessed September 2022.

[52] Dominik Schmidt. simplified-recipes-1m dataset, 2019. Last accessed September 2022.

[53] Niki Segnit. *The Flavor Thesaurus: A Compendium of Pairings, Recipes and Ideas for the Creative Cook.* Bloomsbury USA; Revised edition, 2012.

[54] Muna Saleh Alrazgan Shahad TalalTallab. Exploring the food pairing hypothesis in arab cuisine: A study in computational gastronomy. *Procedia Computer Science*, 82:135–137, 2016.

[55] Nathan Siu. Big food to big data: Do food tv shows influence consumer behavior?, 2019. Published in Medium Website on www.medium.com, Last accessed August 2022.

[56] Songsoptok. Food for thought, 2016. Published in Medium Website on www.medium.com, Last accessed August 2022.

[57] Bruno Stecanella. Understanding tf-id: A simple introduction, 2019. Published in MonkeyLearn Blog, Last accessed September 2022.

[58] Lisa Tagliaferri. An introduction to machine learning, Published 2017, Last Updated 2022. Developer and author at DigitalOcean.

[59] Carolyn E. Tajnai. Samuel was artificial intelligence pioneer. 1991. Last accessed September 2022.

[60] Camellia Tse. Yummly recipe data, 2020. Last accessed July 2022.

[61] A. M. Turing. Computing machinery and intelligence. *Mind*, 49:433–460, 1950.

[62] Madiyar Tyurin. Molecular gastronomy at the world's best restaurants, 2020. Last accessed August 2022.

[63] Nigel Ward. Fried pecorino with honey and summer herbs, .n.d. Published in sbs Website on www.sbs.com.au, Last accessed September 2022.

[64] Rosemarie Willett. Apples cheese: The perfect pair!, 2020. Published in North Tisbury Website on www.northtisburyfarm.com, Last accessed September 2022.

[65] Wanshun Wong. What is a siamese neural network?, 2020. Last accessed July 2022.

[66] Yummly. Yummly recipe data. 2020. Harvard Dataverse, V1, Published on www.dataverse.harvard.edu, Last accessed July 2022.