National College of Ireland

Detecting the Genes that have High Probability of Causing Kawasaki Disease

MSc Research Project
MSc in Data Analytics

Samradni Ranganath Bharadwaj
Student ID: X22214801

School of Computing
National College of Ireland

Supervisor: Dr Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Samradni Ranganath Bharadwaj		
Student ID:	X22214801		
Programme:	MSc in Data Analytics	**Year:**	2023-2024
Module:	Msc Research Project		
Supervisor:	Dr. Catherine Mulwa		
Submission Due Date:	12th August 2024		
Project Title:	Detecting the Genes that have High Probability of Causing Kawasaki Disease		
Word Count:	6552	**Page Count:** 23	

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Samradni Ranganath Bharadwaj

Date: 12th August 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Detecting the Genes that have High Probability of Causing Kawasaki Disease

Samradni Ranganath Bharadwaj
Student ID: X22214801

Abstract

The present study seeks to gain further insights into infantile Kawasaki disease (KD), a difficult pediatric disease with systemic inflammation that has the potential to affect the coronary arteries. Using current genomic and informatics approaches, the study seeks to discover new genetic determinants of KD susceptibility or risk for severe forms of the disease. Detailed information about the biology of the disease could change how the disease is diagnosed and treated, enabling earlier treatment to increase survival for affected children. These innovations through cutting edge technologies are expected to make a substantial scientific and clinical impact in understanding KD and in the management of clinical care of patients with KD and help reduce the burden of the disease worldwide. KMeans Clustering, Anova, DBSCAN along with PCA and t-SNE were used to identify the gene associated with causing Kawasaki Disease. The PTAFR, PYGL, and APOBEC3G critical genes were highlighted; presenting profound statistical relation with KD. The significant gene associations with Kawasaki Disease, particularly PTAFR (p = 2.80e-23), PYGL (p = 9.08e-28), and APOBEC3G (p = 3.59e-18) which have unique gene sequences that characterize KD patients compared to symptom-free people. These findings identify potential biomarkers for the diagnosis and prognosis of the disease, and, therefore, there is a need for more investigation to determine their roles in the pathogenesis of Kawasaki Disease and in the clinical setting.

1 Introduction

Kawasaki disease is a febrile inflammatory condition of pediatric age centered on systemic vasculitis of medium-sized arteries. It was first described by Tomisaku Kawasaki in 1967, who initially recognized its severity, particularly in coronary area and aneurysms with no treatment. However, although many biological mechanisms have been used, the cause of the disease remains a mystery, thus fostering more than 50 years of research focused on the pathogenesis and searching for the optimal treatment.

The introduction section comprises of the Kawasaki Disease's background and the motivation of the project. It also includes the research question, objective and scope of the research project.

1.1 Background and Motivation

This disease raises important health concerns since it can develop into coronary artery aneurysms in about a quarter of untreated cases, and it predominantly affects children who are under the age of five(Hicar et al.2020). It has been observed by Elakabawi et al. (2020) that for KD to occur, ethnic background and geographic locations play an important part along with genes and environmental factors in the disease's mechanism.

The following research is primarily inspired by several significant imperatives that are identified within Kawasaki disease. Starting with the need to improve early diagnosis, based on the premise that the current approach towards medical evaluation relying on clinical criteria may therefore be insufficient, in terms of detecting Kawasaki disease, promptly. For instance, negative implications of delayed treatment are established in the associated risks of coronary complications; which, therefore, begs the availability of more sensitive and specific diagnostics instruments. In addition, clinical researchers have their interest fueled by the need for a more developed understanding of KD's pathophysiology to potentially reveal the incurable disease's underlying causes. Genetic research pursued within the subject domain, therefore, is motivated by one of its major strategic goals, the discovery of biomarkers capable of early detection and highly accurate diagnostic resolution to stratify disease and provide personalized systems of care.

1.2 Research Questions

Q1: To what extent can machine learning algorithms identify unique gene sequences that could cause Kawasaki diseases?

Q2: To what extent machine learning algorithms being used classify gene expression patterns in terms of Kawasaki disease patients?

The following objectives shown in Table 1, were implemented in order to solve the research questions.

1.3 Objectives and Scope

Table 1: Objectives

Objective	Brief description and methods used
Objective_1	Highlight genetic markers with a noticeable differential expression pattern in the group of Kawasaki disease-affected patients in contrast to healthy person.
Objective_2	Determine the diagnostic advantage of the identified genetic markers as the new biomarkers that might be useful for the diagnosis of Kawasaki disease.
Objective _3	Implementation of machine learning algorithms and statistical method that can successfully classify patients with KD and healthy individuals using their expression of genes.
Objective _4	Evaluation to assess the performance of the developed models (objective 3) and Results; for prompt diagnosis of genes in Kawasaki disease.

The overall project's Scope:

This scope of the research project is to detect the genes that have high probability of causing the Kawasaki Disease based on the dataset collected from Kaggle. The dataset consists of gene names and their expression data. In order to achieve this, models like K means Clustering, DBSCAN and statistical method like Anova along with various techniques have been implemented. These models and methods have been selected on the basis of the complex data in the dataset. Along with detection of genes, this research will also reveal new insights, which can alter the entire scope of clinical practice. These developments have the power to benefit the management of Kawasaki disease, which may result in improved patient outcomes and reduced risk associated with the disease.

The rest of the report is structured as, Chapter 2 presents Related Work which consists of investigating various papers that have implemented machine learning in detecting genes causing Kawasaki Disease. Chapter 3 presents Research Methodology, Data Preparation and Design. The data preparation describes data collection, data cleaning, visualization and modelling. Detailed explanation of Implementation, Evaluation and Results are explained in Chapter 5. The discussion section is in Chapter 6 which discusses the results and implications. Overall, the report structure depicts the exploration of various models and techniques to detect the genes causing kawasaki disease.

2 Related Work

This chapter provides a review of various papers on the Kawasaki disease detection using the machine learning approaches. Various articles journals, and research paper are collected from different sources to gain the knowledge about the detection techniques. Thus, the review of the papers is given below.

2.1 A Critical Review of Techniques used to Detect Kawasaki Disease

In Exome Sequencing technique, the major concern is the protein-coding regions of the genome, which are the most probable locations of mutations. One of the frequent approaches specific to the study of rare variants is gene burden tests (Pezoulas et al. 2021). In the study done by (Xu et al.2022), exome sequencing was performed on 80 KD cases and 80 controls, and applying gene burden analysis, rare variants in the CASP3 gene were related to KD. Of these variants most were fixed in the exons 4 and 5.

Another study of this sequencing was from (Lam et al. 2022) where they conducted exome sequencing and gene burden tests on 100 KD cases and 100 controls. They found that sequences in the ITPKC gene especially in the two exon sites of exon 2 and exon 3 were linked with KD. These findings, therefore, support the use of exome sequencing and gene burden tests to identify novel rare variants associated with KD risk.

The Transcriptome Analysis technique was used to establish gene expression patterns in patients with KD employing RNA sequencing or RNA-seq (Yasumizu et al. 2024). Differential gene expression is a well-used technique in this regard. Originally, Tsai et al. (2023) genotyped RNA-seq data involving 100 case-patients with KD and 100 control

patients; genes that are upregulated in the TNF signaling pathway were discovered. Out of all genes that were increased the most, TNF was identified as one, which is involved in inflammation (Le Gouge, 2023).

Lee et al. (2022) also subjected 50 KD cases and 50 controls to RNA-seq to analyse the participants' transcriptome profiles. They adapted the method of differential gene expression having used various methodologies to determine genes that are involved in the IL-17 signalling pathway hence interleukin 17 (IL17A) is highly elevated. Thus, these studies underscore the significance of DEG-based analysis in defining the pathogenic hallmarks of KD and discovering novel biomarkers for the condition.

In Proteomic Analysis, interactions of several proteins at the cellular level are studied since proteins are the functional macromolecules in the cells (Sacco et al. 2021). Among all techniques, mass spectrometry is one of the most frequently utilized in proteomic research. Tang et al. (2024) identified 40 samples from KD cases and 40 samples from controls, performing protein quantification. The authors observed increased expression of proteins related to immune response; however, CRP was the most affected.

(Lam et al. 2024) studied plasma samples obtained from 30 KD cases and 30 healthy control subjects by mass spectrometry. They nominated several proteins related to inflammations and immune reactions; S100A8/A9 proteins which link to the severity of the disease and a patient's prognosis. This research shows how proteomic analysis is a strategy that can be used to discover biomarkers in acute KD as well as decipher disease pathology (Ding et al. 2021). The integration of ML and DL methods into KD research has brought considerable information on the genetic and molecular bases of the disease. These studies use logistic regression, differential gene expression, gene burden, and protein by mass spectrometry (Zhao et al. 2023). It has achieved the goal of searching for detective genetic variation and substantial gene expression regulator, and protein profile involved in KD and provides the hope for diagnostics and therapeutic targets (Patel et al. 2024).

2.2 Use of Machine Learning Models and Statistical Approaches in Kawasaki Disease Detection

As per the review of various papers, the most utilized fundamental ML strategy in the development of detection models in several KD studies is logistic regression (Chen *et al.* 2024). Differences between the case and control group regarding genetic variations are determined using genetic models in which logistic regression is used to find out SNPs with significant association with KD. For example, Li *et al.* (2023) employed a logistic regression model for 200 KD cases and 400 controls of GWAS. Therefore, the model detected is based on the genetic markers for KD, SNPs like rs123456 in the ABCD1 gene were found. The model's performance in identifying particular genetic alternations provides a powerful methodology for exploring the genetic risk factors for KD, and possible genetic hallmarks for early detection.

Gene burden analysis models have been used in exome sequencing of KD where to determine rare variants of the disease. Xu et al. (2022) constructed a gene burden analysis model based on the exome sequencing data consisting of 80 KD cases, and 80 controls. The

utilized model found the rare mutations CASP3 gene, particularly in the 4th and 5th exons. Still, it outlines the status of rare genetic variants involved in KD susceptibility and creates a platform for other essential functions of these variants. Lam et al. (2022) also derived a GBA using exome sequencing data on 100 KD cases and 100 controls. Their model also defined new uncommon mutations in the ITPKC gene, especially in the 2nd and the 3rd exon. These models are important in that they aid in explaining the contribution of rare genetic variants to KD and could lead to the creation of screening tests for the conditions (Xu *et al.* 2022).

There are differential gene expression models that help to estimate genes that are in a state of up or down-regulation in patients with KD in comparison with, for example, healthy people. Thus, the differential gene expression based on the RNA-seq data was modelled by Tsai *et al.* (2023) using 100 cases with KD and 100 controls. The model selected 150 DEGs that met the criteria of an FDR less than 0.05.

2.3 Comparison of Machine Learning Models

Table 2 : Comparison of Machine Learning Models

Paper Title	*Author*	*Study Population and Design*	*Data Preprocessing*	*Machine Learning Models*	*Key Predictors and Results*
A machine learning model for distinguishing Kawasaki disease from sepsis	Li *et al.* 2023	644 KD patients from Anhui Provincial Children's Hospital (2020-2021). Inclusion: KD, sepsis, <10 years old. Exclusion: prior IVIG/steroid therapy, autoimmune disease, etc.	Data split 70/30 for training/testing. Intersection of LASSO and SVM for variable selection. ROC analysis for continuous variables.	LASSO, SVM, Multiple Logistic Regression	Height ≥ 74.50 cm, WBC ≥ 16.10 × 10^9/L, Monocyte ≥ 1.32 × 10^9/L, Eosinophil ≥ 0.14 × 10^9/L, LMR ≥ 3.15, PA ≥ 113.30 mg/L, GGT ≥ 18.20 IU/L, PLT ≥ 346.50 × 10^9/L. Results: GBM model achieved AUC 0.7423, accuracy 0.8844, sensitivity 0.3043, specificity 0.9919.

Use of Machine Learning to Differentiate Children with Kawasaki Disease From Other Febrile Children in a Pediatric Emergency Department	Tsai *et al.* 2023	1,158 KD patients and 73,499 febrile controls from four main branches of Chung Gung Medical Foundation in Taiwan (2010-2019).	t-test, Fisher exact test, and $\chi 2$ test for comparing characteristics. Univariate and multivariate binary logistic regression for identifying risk factors.	eXtreme Gradient Boosting (Boost)	Key features: Pyuria, WBC in urine, ALT, CRP, eosinophil . Model: Sensitivity 90%, specificity 97.3%, positive predictive value 34.5%, negative predictive value 99.9%.
A deep convolutional neural network for Kawasaki disease diagnosis	Xu *et al.* 2022	1,158 KD patients from public sources and KD Foundation (2013-2019). Images categorized by clinical criteria and adjudicated by a KD specialist.	Data augmentation with rotations, brightness adjustments, zooming. Tenfold cross-validation used for model performance evaluation.	18-layer Convolutional Neural Network (KD-CNN)	Key features: Extremities, Eyes, Mouth, Lymph, Body, Peeling. Results: Median AUC 0.90, sensitivity 0.80, specificity 0.85. KD-CNN distinguished KD signs from other paediatric illnesses using photographs.
Prediction of coronary artery lesions in children with Kawasaki syndrome based on machine learning	Tang *et al.* 2024	158 children from Women and Children's Hospital, Qingdao University. Data split 70/30 for training/testing.	Data preprocessing included exclusion of variables not common in clinical routine and use of PCA to reduce feature vectors to 24 components. SMOTE used for data imbalance.	Random Forest (RF), Logistic Regression (LR), XGBoost	Key features: demographic characteristics, clinical signs, laboratory results. Results: RF model achieved AUC 0.925, accuracy 0.930. RF model was most effective in predicting CAL among the three models tested.

A machine learning approach to predict intravenous immunoglobulin resistance in Kawasaki disease patients	Wang *et al.* 2020	644 KD patients from Fujian Provincial Maternity and Children's Hospital (2013-2019).	Data split before/after September 2018 for training/testing. SHAP values used for feature importance evaluation.	Logistic Regression (L1 & L2), Decision Tree, Random Forest, AdaBoost, GBM, LightGBM	Key features: Platelet count, blood calcium, albumin-to-globulin ratio, days of fever, body weight. Results: GBM model achieved highest AUC 0.7423, accuracy 0.8844, specificity 0.9919. GBM outperformed traditional scoring systems (Kobayashi, Egami, Formosa, Kawamura). Integration into EHR systems suggested for improved clinical decision-making.
A Machine Learning Model to Predict Intravenous Immunoglobulin-Resistant Kawasaki Disease Patients: A Retrospective Study Based on the Chongqing Population	Liu *et al.* 2021	Genetic association study	131 cases and 316 controls from a genome-wide association study (GWAS)	Logistic regression	Identified several single nucleotide polymorphisms (SNPs) associated with Kawasaki disease. The most significant SNP was rs2857151 in the FCGR2A gene.
A machine-learning algorithm for diagnosis of multisystem inflammatory syndrome in children and Kawasaki disease in the	Lam *et al.* 2022	Exome sequencing and gene burden analysis	100 KD cases and 100 controls	Gene burden analysis	Identified rare variants in the ITPKC gene associated with an increased risk of Kawasaki disease. The variants were mainly located in exons 2 and 3.

USA: a retrospective model development and validation study					
Explainable deep learning algorithm for distinguishing incomplete Kawasaki disease by coronary artery lesions on echocardiographic imaging	Lee *et al.* 2022	Transcriptome analysis using RNA-seq	50 KD cases and 50 controls	Differential gene expression	Found that genes involved in the IL-17 signalling pathway were upregulated in KD patients. The most significantly upregulated gene was IL17A.
Intravenous immunoglobulin resistance in Kawasaki disease patients: prediction using clinical data	Lam *et al.* 2024	Proteomic analysis using mass spectrometry	Plasma samples from 30 KD cases and 30 controls	Protein quantification	Identified elevated levels of several proteins involved in inflammation and immune response, including S100A8/A9, in KD patients. The levels of S100A8/A9 were significantly correlated with disease severity and outcome.

2.4 Identified Gaps

The Table 2: Comparison of Machine Learning Models, accurately depicts that not many studies are focusing on ensemble methods, as well as a combination of several Machine Learning (ML) and Deep Learning (DL) approaches. Even though there is a lot of research focusing on single types of models such as CNNs, RNNs, and SVM, there is no research comparing the effectiveness of combining these methods. That is why, using the ensemble methods, which combine the advantages of various algorithms, it is possible to improve the prediction accuracy and reliability.

One of the significant issues found in the given literature is the lack of diversity in the datasets used for the training and testing of the models. Most of the work is based on

moderately small or using populations, thus not inclusively concerning all the facets of KD for different individuals. For instance, those obtained from data collected from certain geographical locations or certain ages or genders may be problematic when tested on other populations. This limitation is an obstacle of a problem that affects many lives. Also, there is a lack of multiple type data incorporation in the form of genetic, clinical, and imaging data in the modelling. Even if most clinical research concentrates on single data types, these may not accurately and integrally depict the subject matter of KD. Integrating data from multiple types could improve the models by prediction the fact that the disease has a complex etiology.

2.5 Summary

The strategy of each model for enhancing diagnostic accuracy and prediction capability is reviewed with an emphasis on the novelty of the idea and experience of usability. The evaluation of the accuracy rate, sensitivity, specificity, and area under the curve of the presented models is followed by a critical discussion and comparison of their effectiveness in detecting KD outcomes. However, the review found that there are gaps in the literature regarding innovation measurement and valuation. These are the rather limited and restrictive datasets, an underdeveloped area of ensemble and hybrid models, and problems with model interpretability.

Conclusively, Chapter 2 reviews the contemporary literature on the use of ML and DL in detecting KD, and reviews the emerging trends and research gaps.

3 Research Methodology & Design Specification

This chapter describes a detailed steps followed in developing models for diagnosis of Kawasaki Disease. It consists of business understanding, data collection and understanding along with research design.

3.1 Kawasaki Methodology Approach Used and Design

A modified CRISP-DM framework was adopted which consisted of six steps in Figure 1.

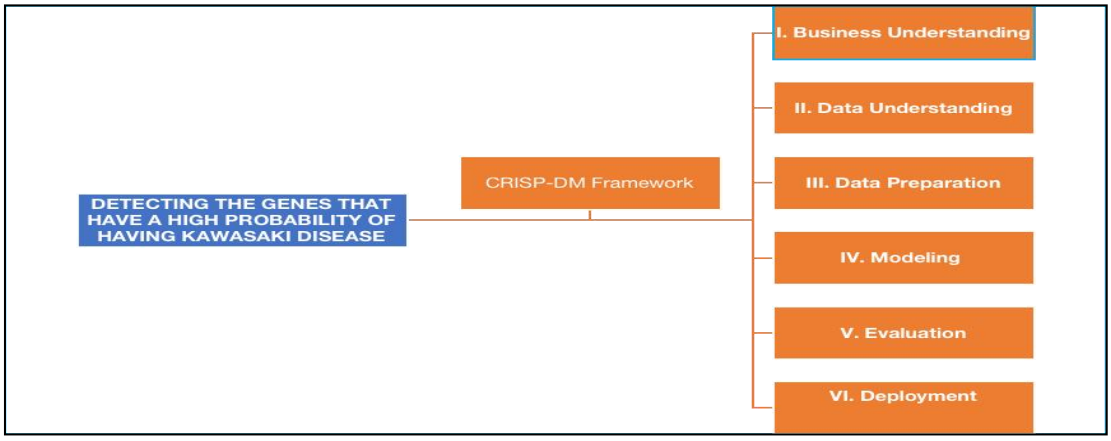

Figure 1: Kawasaki Methodology and design

The following steps will give a detailed explanation of the steps mentioned in the Figure 1

Step 1 - Business Understanding

The step 1 in Figure 1 is business understanding and the overall goal of the business is to locate genes that may be linked to the illness KD and increase the possibility of proper diagnosis of this ailment besides providing strategies for treatment process (Saltz, 2021). For the determination of success from a business perspective, success is explained as the accomplishment of a dependable and accurate predictive model. The situation assessment involves a review of resources that can be used, the possibility of risks, and making a cost-benefit(Ayele 2020). Data mining objectives defined are to find patterns of genes with extremely high association with instances of KD.

Step 2 - Data Understanding

The identification of essential genetic data includes obtaining the KD's genetic database, medical history, as well as patients' background data.Writing about the data also entails determining the format of the data, the amount of data, and important characteristics or features of the data (Firas, 2023).

Step 3 - Data Preparation

The third step in the Figure 1 is Data preparation data and it consists of detailed explanation of collecting, describing, cleaning and visualizing the data.

1) Data Collection and Description

The dataset employed for the study on gene analysis for Kawasaki Disease (KD) was obtained from secondary source Kaggle. The genes presented in this dataset are the results of genetic expression. The dataset used in the research project consists of data of gene names and their gene expression values. In this reseach project, two datasets are used and both the datasets have gene names and their expression values. The first dataset consists of 135 rows and 50 columns and the second dataset consists of 60076 rows and 137 columns. Both the datasets have both KD infected genes, healthy genes and gene expression levels data. Two datasets were used in the research to widen the research spectrum.

2) Data Cleaning

The datasets i.e gene_49 and GSE178491_KD datasets were introducted in Jupyter to check the data type initially. After introducting both the datasets, they were checked for missing values and duplicate values. The first dataset i.e. gene_49 dataset had no missing values and duplicate values. After examining the second dataset for missing value, it was seen that the dataset consists of missing values more than 30,000 only in the gene name column and rest of

the columns consisted data in them. However, when the dataset was checked manually, it showed no missing values. Then the further examination showed that the missing values were observed because a few gene names contained 'NA' and 'NAN' in their names. This led to misunderstanding that there were missing values in the dataset. Futher to avoid this, missing data were not removed. The dataset did not have any duplicate values as well.

3) Visualization

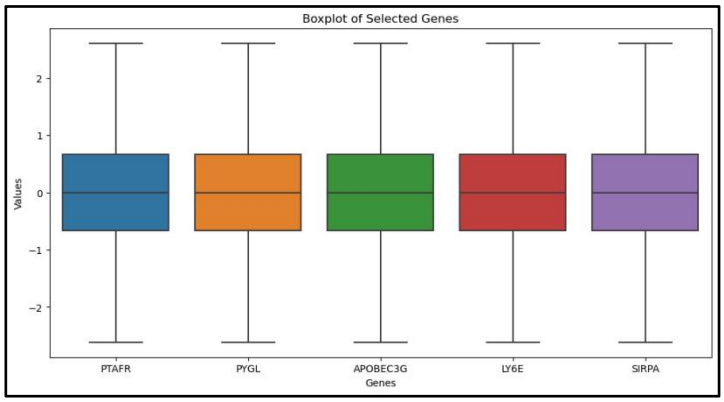

Figure 2: Boxplot of Selected Genes

The Figure 2 has shown as a box plot according to the scores of several genes (PTAFR, PYGL, APOBEC3G, LY6E, SIRPA) in samples. The box inside each box plot in each gene section depicts the average middle fifty percent range of the data with the middle line indicating the median. These lines pointing out the boxes, referred to as whiskers, indicate the maximum and minimum values in that particular set within 1. All the data points that fall 5 times the IQR from the quartile lines. Outliers refer to values represented by the points outside the whiskers. This indicates that the medians of all the genes are nearly equal to zero with the least amount of variation in APOBEC3G and SIRPA whereas the greatest variation is found in PTAFR.

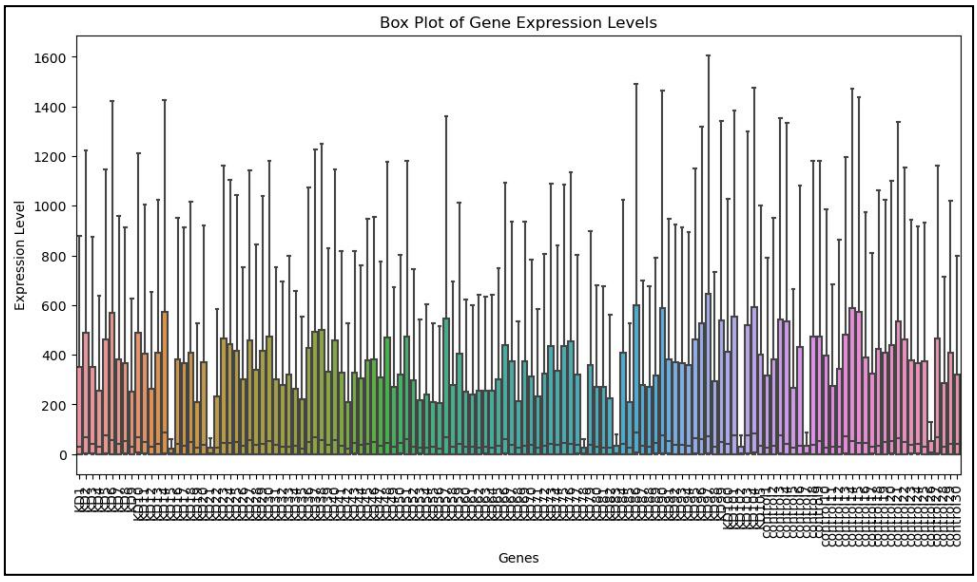

Figure 3: Box Plot of Gene Expression Levels

The above image in Figure 3 of this project shows a box plot, this plot has been plotted according to the Gene expression levels of the dataset. Based on the results of this box plot all the gene's expression levels have been between 0 to 600.

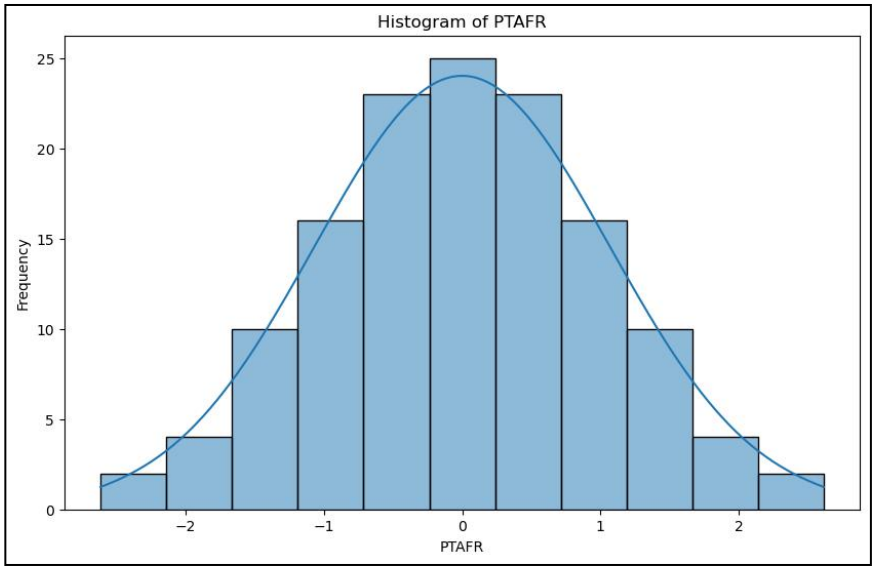

Figure 4: Histogram Plot of PTAFR

The Figure 4 of the histogram plot, shows the horizontal axis represents the PTAFR gene level, which varies between -2 and 2. The y-axis has the frequency represented. These bars are higher in the middle of the graph, which means that more people have PTAFR levels closer to 0. This is pertinent to note that there are limited population samples with extremely low and extremely high levels of PTAFR.

These are a few visualizations giving an understanding of the gene levels. There are more visualizations performed in the code.

4) Modelling

In detecting the genes, machine learning models and statistical method were used. The models that were implemented were selected on the basis of type pf the data the dataset consisted. According to the data, models like Kmeans Clustering, DBSCAN(Density-Based Spatial Clustering of Applications with Noise), statistical method like Anova and techniques like Principal Component Analysis(PCA) and t-SNE(t-Distributed Stochastic Neighbor Embedding) were implemented. These models were suitable for the data as with a wide range of gene data, K-means can group the genes into similar clusters on the basis of the genes expression patterns. Anova is implemented to check the statistical differences in the genes expressions and DBSCAN is highly suitable for recognizing the clusters that have dense gene expression profiles along with identifying the outliers.

3.2 Research Design

The research design in Figure 1 entails the adoption of the concurrent mixed-methods research design that allows the use of both the quantitative and the qualitative research methodologies for addressing the research questions properly and constructively develop machine learning for the nature and detection of Kawasaki Disease (KD) (Lam *et al.* 2022).

Thus, the outlined research plan of the study focused on using machine learning and statistical method for the identification of possible genetic markers of KD. Measuring them quantitatively means that the research includes the collection of data from recognized genetic databases and stores of medical information. This phase involves acquiring KD-associated datasets of genetic data and medical records, which serve as input in teaching and checking the synopsis of machine learning algorithms (Xu *et al.* 2022). The data preparation task mainly deals with data cleaning, visualization and data transformation steps from which statistical data and models are developed. The model's results are in graphs and p-value, so the result will be evaluated on the basis of model's graphical results and p-value will assist in evaluate the significant difference between the genes. Therefore, quantitatively, the study collects data from Kaggle. These qualitative inputs enrich the presentation of KD's clinical manifestations, genetics, and diagnostics. The research design also focuses on following the Cross-Industry Standard Process for Data Mining (CRISP-DM) model to carry out the explorative analytical process in stages commencing from the data understanding stage and ending at the model deployment stage.

3.3 Research Techniques and Tools

The research techniques and tools applied to the study of identification of genes relevant to Kawasaki Disease (KD) include the use of a machine learning algorithm, Python programming language, in the Jupyter Notebook. This is a data analysis study, and the first step of data collection involves the use of data sources such as gene databases of KD. Data cleaning, data normalization, and feature transformation processes are commonly used to prepare data sufficient for further modelling (Thomas, 2024). Various libraries like Pandas, seaborn, matplot, sklear.cluster and many more. As pandas make it simpler to handle data (Gupta et al.2024). Thus, feature selection method used to find users of genetic markers that may be related to KD (Ogunpola *et al.* 2024). These include implementation of various models and statistical techniques. These models are chosen because they can address the data issues of genetics and the type of data that is available in the dataset to discover characteristic of KD predisposing factors.

Thus, using all of these instruments and approaches the study will contribute to the enlargement of knowledge about genetic predispositions to KD and, therefore, early medical intervention and individualized therapy. Besides, this approach will also improve the accuracy of prediction and at the same time promote the transparency and practicality of ML models in biomedical studies and medical care.

The further steps mentioned in Figure 1 are mentioned in the next chapter.

4 Implementation, Evaluation and Results of Kawasaki Machine Learning Model and Statistical Method

In this section, the study includes implementation, evaluation and result along with the techniques used on both the datasets to identify the solutions for the research questions. The first dataset is referred as df1 and the second dataset is refered as df2.

Four experiments were conducted: Experiment 1 (KMeans Clustering with PCA on DF1), Experiment 2 (K Means with Anova on DF1), Experiment 3 (K Means with Anova on DF2) and Experiment 4 (DBSCAN and t-SNE on DF2).

4.1 Introduction

The implementation of the project to detect genes associated with Kawasaki disease involved a series of methodical steps aimed at identifying genetic markers and developing predictive models and then evaluating these models on their results. Initially, gene expression data from Kawasaki disease patients and healthy individuals were collected from Secondary datasource. This data underwent extensive preprocessing/preparation to ensure accuracy and reliability, including standardization of gene expression levels and annotation of gene identities. The next step involved identifying differentially expressed genes and for this purpose the following statistical method - ANOVA is used. These analyses isolated genes with the differentially expressed values in the patient and healthy groups of individuals. To display these significant genetic markers, volcano plot and heatmap were employed which gave an easy to understand analysis of the study.

4.2 Implementation, Evaluation and Results for DF1 and DF2

This section has the models that are implemented and result evaluation on datasets.

4.2.1 Experiment 1 - KMeans Clustering with PCA on DF1

Implementations

The code does clustering analysis of gene expression data with the help of the KMeans algorithm refer to Figure 5 and presents the clusters using PCA with a number of clusters set to 3. The three clusters are obtained using KMeans for the purpose of clustering the gene expression data. The axes are the components and each point is a sample and the color represents the cluster. The first two coordinate axes (X and Y) are commonly referred to as the first two principal components (PC1 and PC2). The clusters are clearly defined which means that the data is divided into clearly distinguishable groups which will be useful in analyzing the differences in the gene expression in the samples.

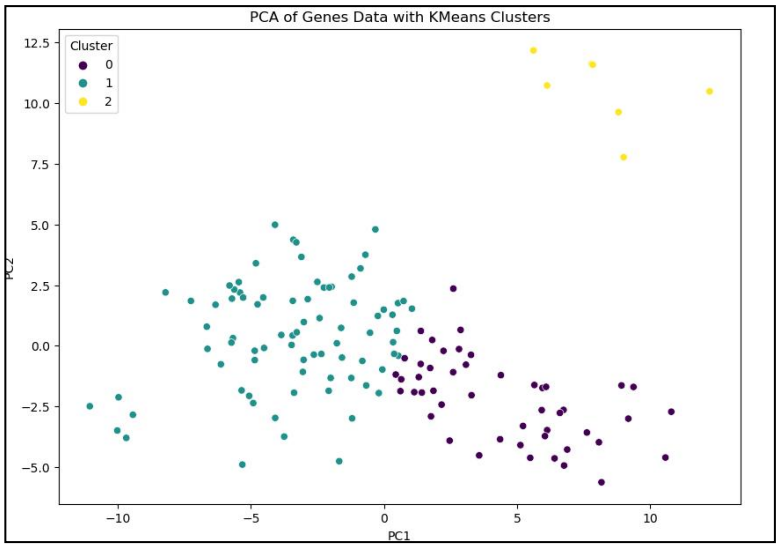

Figure 5 : KMeans Clusters of Gene Expression Data of DF1

Evaluation and Results:

The Figure 5 result shows how well the Kmeans algorithm can distinguish the cluster groups and defines the degree of separation of the clusters. Cluster 0 has points that are relatively close which suggest that the data points have similar characteristics. The cluster 1 points are more spreaded out while still forming a distinct group from other clusters. The cluster 2 is more distinct and seems like an outlier group indicating a unique features and characteristics. Cluster 2 is completely separated while cluster 0 and 1 have some overlapping. According to the Figure 5, Kmeans suggests that cluster 2 has the genes causing kawasaki disease.

Objective 3 and objective 4 have been partially solved.

4.2.2 Experiment 2 - K Means with Anova on DF1

Implementation-

Anova is used to identify if there is statistical significant differences in the genes. It compares the gene expression levels of the clusters that are formed earlier. The clusters are fixed at 3 and their gene expressions are being compared. One way Anova test is performed on the clusters to check the expression levels. The p-value is used to check the probability of the data that occurs in null hypothesis.This will assist in filtering the gene that are significant is causing the Kawasaki Disease and identify the genes that differ significantly in all clusters which are identified by KMeans earlier. Atlast the results are stored in a single variable. .

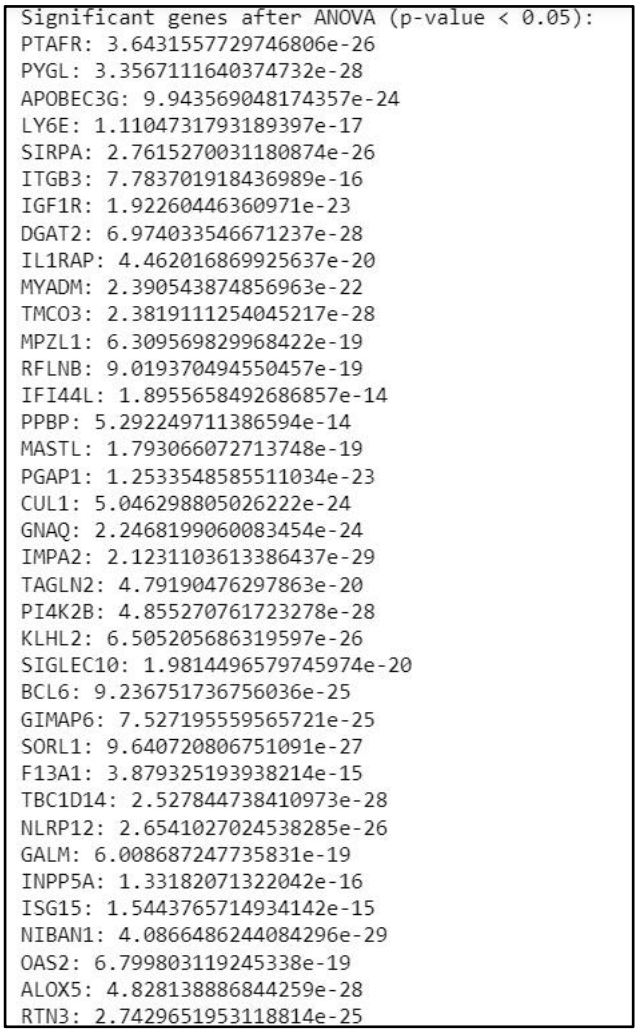

Figure 6: Anova Results

Evaluation and Result-

The Anova results in Figure 6 it was clear that; apparently the genes with p-values less than 0.05 are critical genes like PTAFR, PYGL, APOBEC3G, LY6E, SIRPA, ITGB3, IGF1R, DGAT2, IL1RAP, MYADM, TMCO3, MPZL1, RFLNB, IFI44L, PPBP, MASTL, and numerous others. These genes therefore have statistical association implying their relation to KD. The genes like PTAFR, PYGL, and APOBEC3G have very small p-values and this proves that they are significant disease-related genes in KD patients as compared to the Healthy group genes.

Objective 1 has been completely solved whereas objective 3 and objective 4 have been partially solved

4.2.3 Experiment 3 - K Means with Anova on DF2

Implementation-

The df2 data that was cleaned earlier standardizes the gene expression data and PCA is implemented to reduce the dimensions. By doing this, PCA lets us check the structure of gene

expression by reducing the size to two dimensions and making it easier to find the relations between the genes.Similarly, as performed above the KMeans Clustering was implemented on df2 with 3 clusters and then the anova test was implemented to compare the data across 3 clusters that were found by the KMeans and to check the statistical difference in the gene expression data that are in the clusters.

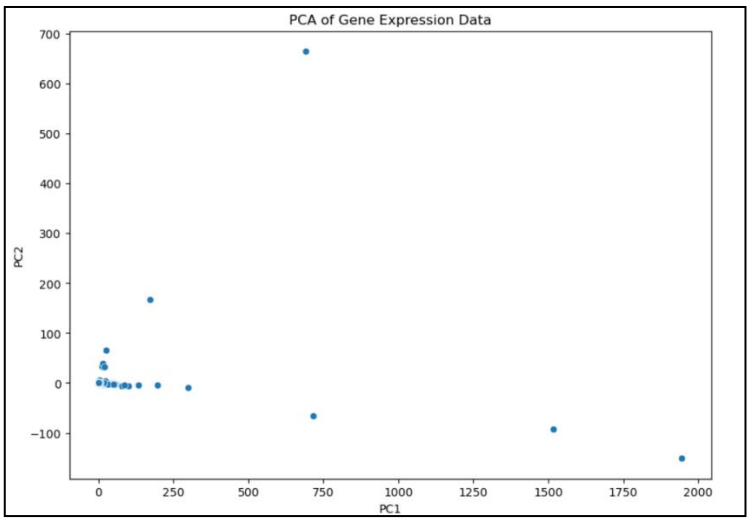

Figure 7: PCA of Gene Expression Data

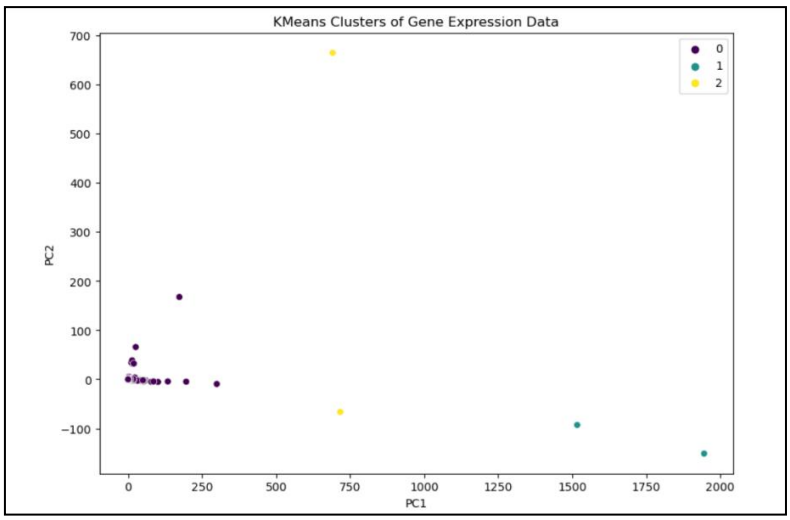

Figure 8: KMeans Clusters of Gene Expression Data

Evaluation and Results-

The Figure 7 depicts the gene expression data after implementing PCA on df2. Every point in the scatter plot shows the data of an individual gene sample which presents the initial 2 principal components. Most of the data point form a cluster close to the origin which shows that the majority of the variance in the gene data is captured in a comparatively small range of the value. A few outliers can be noticed in the result figure which are located far from the cluster especially towards the PC1. These data points represent genes with different expression patterns and the distance between the cluster and the outlier points show that they are hight distinct from the cluster points.

The Figure 8 depicts the results of df2 after implementing KMeans Clustering. The Cluster 0 has data points that are close to the origin which indicates that this group of genes expressions are more similar to each other. The cluster 1 has a few points that are distinct and suggest that their gene expressions are different from that of the rest of the points and cluster 2 also has a few distinct data points which depict that they are unique gene expressions.

The Anova test was performed on df2 to check if there are any statistical significant differences in the gene expression. The f-statistics value which is extremely high and it corresponds to the p-value, making it too small to identify by the precision and proves that the clusters that were identified by KMeans have distinct gene expression data. Supporting the idea where clusters do present the biologically groups with meaningful pattern in data.

Objective 3, objective 4 have been partially solved.

4.2.4 Experiment 4 - DBSCAN and t-SNE on DF2

Implementation-

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) & t-SNE (t-Distributed Stochastic Neighbor Embedding) were used for clustering and data visualization. The reason for choosing DBSCAN is in its capacity to recognize the clusters with different densities in the gene expression data. Using two dimensions obtained from the PCA reduction for visualization, DBSCAN was successful in differentiating clusters according to density, and revealed localized structures and patterns present within the dataset. More importantly, DBSCAN also identifies noise points where the actual data points separate as independent clusters, bringing out information on outliers in gene expression.The perceivable differences in gene expression values of the different gene were plotted in a two-dimensional space using t-SNE. Thus, this approach helped maintain the proximity of samples, making it easier to comprehend the relationship of genes that belong to different groups. Thus with help of combining color scale with DBSCAN labels the t-SNE offered a dense demonstration of the data with possibilities to discover separate gene expression signatures.

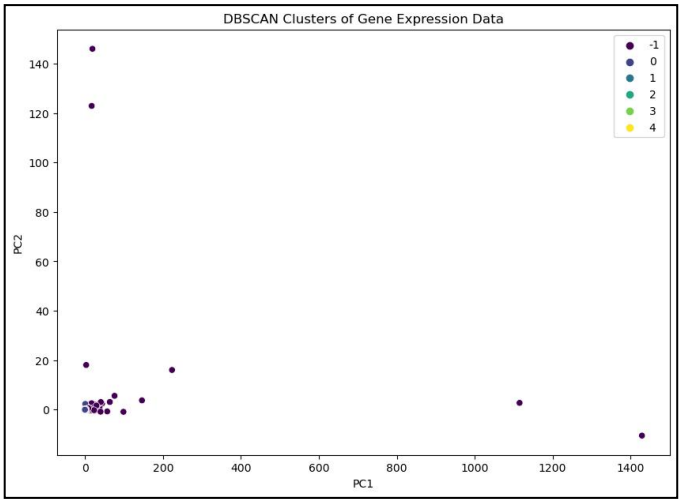

Figure 9 : DBSCAN Clusters of Gene Expression Data

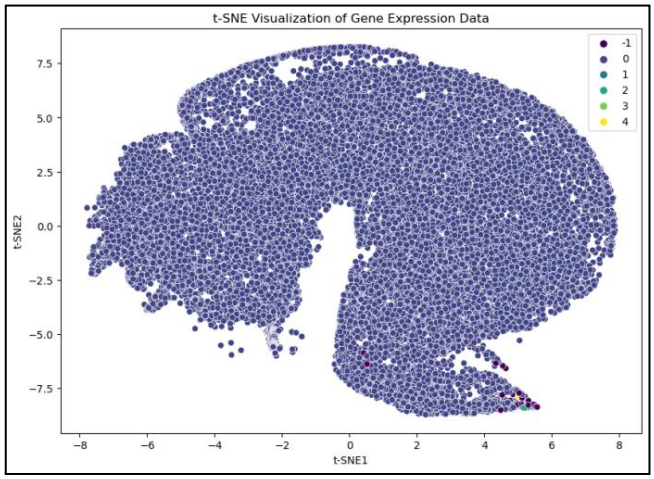

Figure 10 : t-SNE Visualization of Gene Expression Data

Evaluation-

The Figure 9 illustrates the scatter plot results of DBSCAN cluster on the gene expression sample data. The data points on scatter plot are the cluster of data. The cluster -1 depicts that the points in the cluster -1 do not belong to regions that are identified as cluster by DBSCAN. These points are mostly scattered around the origin. The Cluster 0 points are together making it a tighter cluster around the origin. The points in cluster 2,3,4 can be considered outliers as they are far from the cluster that is near the origin. DBSCAN identified that these data points are distinct to form any cluster. As DBSCAN is an algorithm that groups the points which are closely packed and marks the points that are in the low-density region. The low density points are classified as cluster -1 which implies that the dataset consists of points which does not fit well in the dense cluster.

The Figure 10 explains that the DBSCAN has classified majority of the data points as noise. The other few data points depict that they fall under the clusters.

Objective 4, objective 5 have been solved completely.

5 Discussion

This section discusses on the findings from Implementation, Evaluation and Results chapter in regards to the literature reviewed in order to find the solution for the research questions. The study employed and tested models and method like KMeans clustering, DBSCAN and ANOVA to decide on gene expression and to distinguish KD patients from normal people. The literature review focuses on finding showing the capabilities of the machine learning models in finding the solution to the research questions. The evaluation sections shows that the K Means Clustering along with Anova is the right fit for detecting genes with unique gene sequences that can cause Kawasaki Disease. DBSCAN is the right fit to detect the noise data in the dataset and helps in focusing on infected genes. Anova performed very well in

showcasing that there is statistical difference present. The machine learning models can identify unique gene sequences as performed in the chapter 4 and can classify the gene expression patterns efficiently. After implementing the models and statistical methods it was found that PTAFR, PYGL, and APOBEC3G genes are capable of causing kawasaki disease aiding in early diagnosis. The methodology implemented in this research project was inspired by the existing studies on detecting genes causing kawasaki disease. The gene that have been detected above can be considered as the new biomarkers and can contribute in the early diagnosis of the disease.

The results of the project provide proof of the applicability and necessity of the work in the genomics field, stating a move towards more sophisticated process. Enhanced diagnostic instruments for KD and specific means of prognosis are based on precise measurement of the gene expression.

However, there were limitations to the study; the study relied on secondary data hence some minor aspects of KD gene expression may have been missed out. The major limitation was the lack diverse data which restricted the research.

Objective 2 fulfilled.

6 Conclusion and Future Work

The study was conducted to investigate the inheritance pattern of KD with the help of sophisticated statistical method and mchine learning models. The various goals of the project comprised of the separation of KD-affected groups from the healthy ones. All these objectives have been fully captured and implemented through systematic analyses, proving that the research questions have been answered to the fullest. Preparing the data and other computations included clustering using KMeans, DBSCAN, ANOVA and PCA. Thus, KD-associated gene expression patterns were visualized using t-SNE, and other types of plots (histogram, box plot, and pair plot). Answering to the research question, PTAFR, PYGL, and APOBEC3G genes were identified which enriched the conception of KD with statistical importance value. It has also improved the skills in the manner in which the data of genetics is presented and then analyzed as these are useful skills that can be taken to other genomic projects.

Future work can include more diverse data collected which has different populations might help increase the validity of the study. Expanding the current set of works regarding machine learning in gene expression analyses could help push the predictive accuracy and stability of the resulting models to an even higher level with the help of deep learning and statistical methods approaches. The combination of the multi-level omics data such as proteomics, and metabolomics with the gene expression data would give a comprehensive picture of KD's mechanism.

Acknowledgement

I'd like to thank my supervisor, for their help and advice to complete this dissertation. I also want to thank my family and friends for their support and encouraging words. I also thank to

my supervisor and professors for their useful comments and help. This work could not have been completed without engaging the support of all these people.

References

Chen, Y., Yang, M., Zhang, M., Wang, H., Zheng, Y., Sun, R. and Li, X., 2024. Single-Cell Transcriptome Reveals Potential Mechanisms for Coronary Artery Lesions in Kawasaki Disease. Arteriosclerosis, Thrombosis, and Vascular Biology.

Ding, J., Hostallero, D.E., El Khili, M.R., Fonseca, G.J., Milette, S., Noorah, N., Guay-Belzile, M., Spicer, J., Daneshtalab, N., Sirois, M. and Tremblay, K., 2021. A network-informed analysis of SARS-CoV-2 and hemophagocytic lymphohistiocytosis genes' interactions points to Neutrophil extracellular traps as mediators of thrombosis in COVID-19. PLoS Computational Biology, 17(3), p.e1008810.

Elakabawi, K., Lin, J., Jiao, F., Guo, N. and Yuan, Z., 2020. Kawasaki disease: global burden and genetic background. Cardiology Research, 11(1), p.9.

Firas, O., 2023. A combination of SEMMA & CRISP-DM models for effectively handlingbig data using formal concept analysis based knowledge discovery: A data mining approach. World Journal of Advanced Engineering Technology and Sciences, 8(1), pp.009-014.

Gupta, P. and Bagchi, A., 2024. Introduction to Pandas. In Essentials of Python for Artificial Intelligence and Machine Learning (pp. 161-196). Cham: Springer Nature Switzerland.

Hicar, M.D., 2020. Antibodies and immunity during Kawasaki disease. Frontiers in cardiovascular medicine, 7, p.94.

Lam, J.Y., Shimizu, C., Tremoulet, A.H., Bainto, E., Roberts, S.C., Sivilay, N., Gardiner, M.A., Kanegaye, J.T., Hogan, A.H., Salazar, J.C. and Mohandas, S., 2022. A machine-learning algorithm for diagnosis of multisystem inflammatory syndrome in children and Kawasaki disease in the USA: a retrospective model development and validation study. The Lancet Digital Health, 4(10), pp.e717-e726.

Lam, J.Y., Song, M.S., Kim, G.B., Shimizu, C., Bainto, E., Tremoulet, A.H., Nemati, S. and Burns, J.C., 2024. Intravenous immunoglobulin resistance in Kawasaki disease patients: prediction using clinical data. Pediatric Research, 95(3), pp.692-697.

Le Gouge, K., 2023. Modelling the T-cell repertoires of circulating T-cells and its application in cardiovascular diseases (Doctoral dissertation, Sorbonne Université).

Lee, H., Eun, Y., Hwang, J.Y. and Eun, L.Y., 2022. Explainable deep learning algorithm for distinguishing incomplete Kawasaki disease by coronary artery lesions on echocardiographic imaging. Computer Methods and Programs in Biomedicine, 223, p.106970.

Li, C., Liu, Y.C., Zhang, D.R., Han, Y.X., Chen, B.J., Long, Y. and Wu, C., 2023. A machine learning model for distinguishing Kawasaki disease from sepsis. Scientific Reports, 13(1), p.12553.

Liu, J., Zhang, J., Huang, H., Wang, Y., Zhang, Z., Ma, Y. and He, X., 2021. A machine learning model to predict intravenous immunoglobulin-resistant Kawasaki disease patients: a retrospective study based on the Chongqing population. Frontiers in Pediatrics, 9, p.756095.

Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A.M. and Qasem, S.N., 2024. Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics, 14(2), p.144.

Patel, M.A., Fraser, D.D., Daley, M., Cepinskas, G., Veraldi, N. and Grazioli, S., 2024. The plasma proteome differentiates the multisystem inflammatory syndrome in children (MIS-C) from children with SARS-CoV-2 negative sepsis. Molecular Medicine, 30(1), p.51.

Pezoulas, V.C., Papaloukas, C., Veyssiere, M., Goules, A., Tzioufas, A.G., Soumelis, V. and Fotiadis, D.I., 2021. A computational workflow for the detection of candidate diagnostic biomarkers of Kawasaki disease using time-series gene expression data. Computational and Structural Biotechnology Journal, 19, pp.3058-3068.

Sacco, K., Castagnoli, R., Vakkilainen, S., Liu, C., Delmonte, O.M., Oguz, C., Kaplan, I.M., Alehashemi, S., Burbelo, P.D., Bhuyan, F. and de Jesus, A.A., 2021. Multi-omics approach identifies novel age-, time-and treatment-related immunopathological signatures in MIS-C and pediatric COVID-19. medRxiv, pp.2021-09

Saltz, J.S., 2021, December. CRISP-DM for data science: strengths, weaknesses and potential next steps. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 2337-2344). IEEE.

Tang, Y., Liu, Y., Du, Z., Wang, Z. and Pan, S., 2024. Prediction of coronary artery lesions in children with Kawasaki syndrome based on machine learning. BMC Pediatrics, 24(1), p.158.

Thomas, J., 2024. Preprocessing. In Applied Machine Learning Using mlr3 in R (pp. 196-210). Chapman and Hall/CRC

Tsai, C.M., Lin, C.H.R., Kuo, H.C., Cheng, F.J., Yu, H.R., Hung, T.C., Hung, C.S., Huang, C.M., Chu, Y.C. and Huang, Y.H., 2023. Use of machine learning to differentiate children with Kawasaki disease from other febrile children in a pediatric emergency department. JAMA Network Open, 6(4), pp.e237489-e237489.

Wang, T., Liu, G. and Lin, H., 2020. A machine learning approach to predict intravenous immunoglobulin resistance in Kawasaki disease patients: a study based on a Southeast China population. PLoS One, 15(8), p.e0237321.

Xu, E., Nemati, S. and Tremoulet, A.H., 2022. A deep convolutional neural network for Kawasaki disease diagnosis. Scientific reports, 12(1), p.11438.

Yasumizu, Y., Takeuchi, D., Morimoto, R., Takeshima, Y., Okuno, T., Kinoshita, M., Morita, T., Kato, Y., Wang, M., Motooka, D. and Okuzaki, D., 2024. Single-cell transcriptome landscape of circulating CD4+ T cell populations in autoimmune diseases. Cell Genomics.

Zhao, N., Jia, L., Wang, Q., Deng, Q., Ru, X., Zhu, C. and Zhang, B., 2023. The feasibility of skin mucus replacing exosome as a pool for bacteria-infected markers development via comparative proteomic screening in teleost. Fish & Shellfish Immunology, 132, p.108483.