MACHINE LEARNING IN THE IDENTIFICATION OF OBESITY IN CHILDREN USING THEIR MOLECULAR DATA

UNDERGRADUATE RESEARCH PROPOSAL SUBMITTED

IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE DEGREE OF

BACHELOR OF THE SCIENCE OF ENGINEERING

Submitted by:

SUBASINGHE S.A.B.D. (2019/E/136)

SUBASHKAR S.(2019/E/137)

DEPARTMENT OF COMPUTER ENGINEERING

FACULTY OF ENGINEERING

UNIVERSITY OF JAFFNA

[MAY] [2023]

SIMPLE OBESITY PREDICTION TECHNIQUE IN CHILDERN BY MACHINE LEARNING

Supervisor: Dr. P. Jeyananthan

Examination Committee:

Lecturer 1 .....

Lecturer 2 .....

CONTRIBUTION TO THE PROPOSAL BY THE MEMBERS IN GROUP:

Sections	2019/E/136	2019/E/137
CHAPTER 1: INTRODUCTION		
1.1 Motivation and Overview	✓	
1.2 Aims and Objectives	✓	✓
1.3 Research Scope	✓	
CHAPTER 2: LITERATURE REWIEW		
2.1 Introduction	✓	✓
2.2 Forecasting Models	✓	
2.2.1 Forecasting Models	✓	✓
2.3 Performance Analysis	✓	
2.4 Available Databases	✓	
CHAPTER 3: METHODOLOGY AND RESEARCH PLAN		
3.1 Data Collection	✓	✓
3.2 Data Preprocessing	✓	✓
3.3 Differential Methylation Analysis		✓
3.4 Clustering and Visualization		✓
3.4.1 Clustering Samples		✓
3.4.2 Clustering CPG sites		✓
3.4.3 Heatmaps		✓
3.4.4 Scatter Plots		✓
3.4.5 Principal Component Analysis		✓
3.5 Functional Enrichment Analysis		✓
3.6 Choose Machine Learning Algorithm	✓	✓
3.6.1 Train the Model	✓	✓
3.6.2 Evaluate Model Performance	✓	✓
3.7 Model optimization and Testing	✓	
3.8 Time Line		
CHAPTER 4: PROGRESS TO DATE		
4.1 Literature Review	✓	✓
4.2 Database Collection		✓
4.2.1 Dataset Selection		✓
4.3 Database Preparation		✓
REFERENCE	✓	✓

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION

1.1 Motivation and Overview

1.2 Aims and Objectives

1.3 Research Scope

CHAPTER 2: LITERATURE REWIEW

2.1 Introduction

2.2 Forecasting Models

2.3 Performance Analysis

2.4 Available Databases

CHAPTER 3: METHODOLOGY AND RESEARCH PLAN

3.1 Data Collection

3.2 Data preprocessing

3.3 Differential Methylation Analysis

3.4 Clustering and Visualization

CHAPTER 4: PROGRESS TO DATE

4.1 Literature Review

4.2 Database Collection

    4.2.1 Dataset Selection

4.3 Database Preparation

REFERENCE

ABBREVIATIONS AND ACRONYMS

WHO	World Health Organization
LR	Logistic Regression
ML	Machine Learning
NB	Naive Bayes
SLR	Systematic Literature Review
ANNs	Artificial Neural Networks
BMI	Body Mass Index
RNN	Recurrent Neural Network
CV	Cardiovascular
GB	Gradient Boosting
CHD	Coronary Heart Disease
RQ	Research Questions
AI	Artificial Intelligence
SVM	Support Vector Machine
RM	Regression Model
DT	Decision Tree
KNN	K-Nearest Neighbors RF Random Forest

Chapter 1: INTRODUCTION

1.1Motivation and Overview

Obesity and its attendant conditions have become major health problems worldwide, and obesity is currently ranked as the fifth most common leading cause of death globally [2]. Researchers have estimated that numbers of adult obesity patients have reached 641 million in 2014 compared to only 105 million in 1975 [2] and obesity causes around three million death per year, thus showing an alarming increase of obesity throughout the world [1].

The worldwide prevalence of childhood overweight and obesity have increased from 4.2% in 1990 to 6.7% in 2010 and this trend is expected to reach 9.1% in 2020. This global health problem is gradually affecting each continent of the world [3].

The World Health Organization (WHO) defines obesity as an "abnormal or excessive fat accumulation that may impair health," further clarifying that "the fundamental cause of obesity and overweight is an energy imbalance between calories consumed and calories expended" [2]. The unit of "Body Mass Index" (BMI), which is measured by calculating [(weight in kg)/(height in m2)], is a simple index intended to classify adults into one of three categories: "underweight," "overweight," or "obese."

The WHO classifies adult obesity using certain BMI cutoffs (Table 1). This WHO classification is beneficial in distinguishing individuals who may be at increased risk of morbidity and mortality due to obesity [2].

BMI value	Obesity Condition
<18	Underweight
18.5 – 24.9	Normal weight
25 – 29.9	Overweight
30 - 34.9	Class I Obesity
35 – 39.9	Class II Obesity
>40	Class III Obesity

Table 1

Multiple studies have demonstrated that obesity is not a simple problem but a complex health issue emerging from a combination of individual factors and substantial causes (unhealthy societal or cultural eating habits) [2]. Most researchers also agree that obesity is an acquired disease that heavily depends on lifestyle factors such as low rates of physical activity and chronic overeating, despite its genetic and epigenetic influences. Mainly effecting factors are:

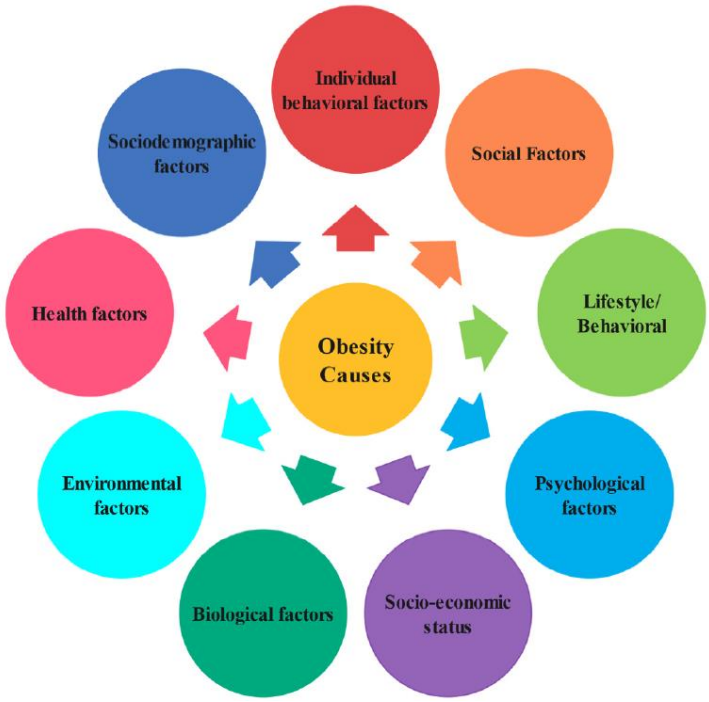


Figure 1: Obesity Causes

Related diseases:

Researchers have also noted that various forms of obesity are based on the area of fat deposition, including peripheral obesity, central obesity and endocrine obesity.

- Peripheral: Accumulation of excess fat in the hips, buttocks and thighs.
- Central: Accumulation of excess fat in the abdominal area.
- Endocrine: Caused by hormonal imbalances such as hypothyroidism and hypercortisolism.

The dangerous profile of obesity is related to increased risk of several chronic conditions and diseases, which include asthma, cancer, diabetes, hypercholesterolemia and cardiovascular diseases. Thus, while obesity is

undoubtedly a condition, it also exacerbates pre-existing conditions and instigates new ones [2].

Obesity can affect nearly every organ system, from the cardiovascular (CV) system to the endocrine system, central nervous system, and the gastrointestinal (GI) system. In addition, obesity is associated with the growing prevalence of several CV conditions, from hypertension and coronary heart disease (CHD) to atrial fibrillation (AF) and even total heart failure [2].

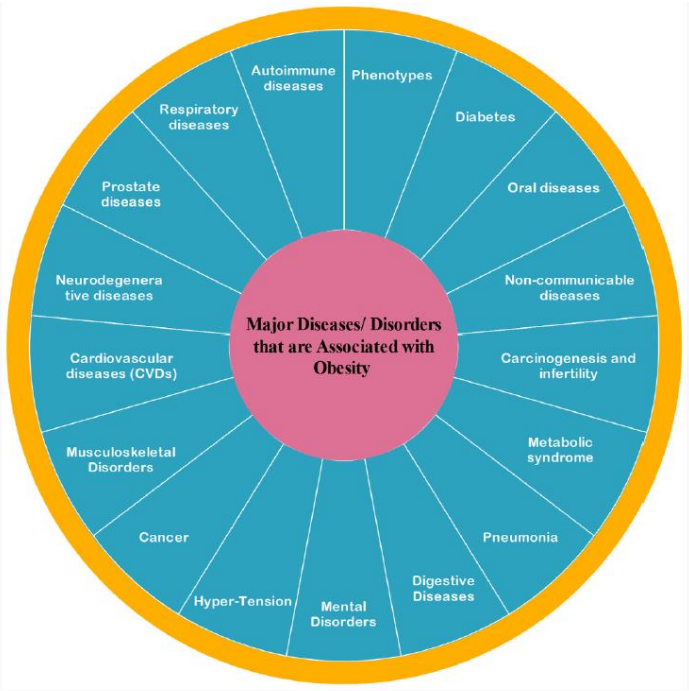


Figure 2: Obesity Associated Diseases

While previously uncommon in young children, obesity is now a worldwide epidemic affecting over 40 million children under the age of 5 [1,2]. Obesity in childhood is associated with both adverse outcomes like hyperlipidemia, diabetes and hypertension, as well as with higher morbidity and mortality in adulthood. The underlying causes of obesity are modifiable risk factors throughout the life course; these risk factors represent major causes of health inequalities. Thus, the prevention of obesity is considered a national and global health priority [5].

Machine learning methods can be classified as supervised learning and unsupervised learning. Supervised learning uses labeled data and tries to predict the outcomes from the input variables. Unsupervised learning uses unlabeled data

and aims to find hidden relationships between variables. Both methods are fast and efficient, thus have been widely applied in many fields such as healthcare, finance and autonomous cars. The purpose of using machine learning is diverse, including but not limited to, extract useful information from data, recognize hidden patterns, acquire knowledge, predict the future and make recommendations. Applications of machine learning have enriched traditional data analytic methods in various fields including health [1].

1.2 Aims and Objectives

Aim:

The aim of this research is to leverage machine learning techniques and molecular data analysis to develop a more accurate and predictive approach for identifying simple obesity in children.

Goal:

The primary goal of this research is to develop a robust machine learning model that can accurately classify children as either obese or non-obese using their molecular profiles. By achieving this goal, we aim to contribute to the field of obesity research by providing a more precise and data-driven approach for identifying simple obesity in children. Additionally, the research aims to explore and identify novel biomarkers and gain deeper insights into the molecular basis of obesity, ultimately contributing to the development of personalized approaches for prevention and intervention.

1.3 Research Scope

The research focuses on utilizing molecular data, including gene expression profiles, epigenetic markers, and other relevant molecular information, to investigate the underlying biological processes associated with obesity in children. The scope encompasses the development and application of machine learning algorithms to effectively classify children as either obese or non-obese based on their molecular profiles.

Chapter 2: LITERATURE VIEW

2.1 Introduction

Nowadays, obesity has become one of the leading causes of death. Overweight and obesity are primary risk factors for many chronic diseases and health conditions, including cardiovascular diseases, type 2 diabetes (T2D), hypertension, and cancers in adulthood. Accordingly, acting to reduce teen obesity can also reduce adult obesity. Early action is one of the most suitable approaches because once children have become overweight, this trend often exists through their adolescence and adulthood [1].

We have gone through some review papers on obesity which searched for machine learning techniques can be used to identify obesity.

Further we have selected some research papers on molecular side of obesity to understand the connections between molecular pathways and obesity. There were different types of genetic factors that have been associated with obesity [2]. Mendelian forms of obesity (single gene disorder), Recessive forms of monogenic obesity (single gene disorder), Partial gene deficiency, genome structural variants, Mendelian Forms of obesity.

Then we have gone through research papers

Datasets we have searched, relevant to our project.

- Genomic
- Transcriptomic
- Proteomic
- Metabolomic
- Epigenetic

In our research we are focused on methylation data which was relevant to Epigenetic data.

2.2 Forecasting Models

Logistic regression: is used to test a predefined hypothesis and find a relationship between input and output variables when the output variables are categorical in nature (i.e., weight gain or loss). Linear regression is similar to logistic regression in terms of examining the association between input and output variables. Its output is continuous, not binary variable. It also assumes a linear relationship between input and output variables [5] [6].

Decision tree: is an algorithm that uses functions to classify data in a shape that is similar to a tree structure. It classifies data by sorting them from root down to leaf node. Each node in the tree represents a variable and each branch from a node represent a possible value of the attribute. It is applied to classify samples to specific classes based on their values. These classes are divided based on specific calculated thresholds. Decision tree is simple, easy to apply, uses both categorical and numerical data and produce promising results [3] [7] [8].

Random forest: is basically a large number of decision trees, could be a couple of hundreds that would function in aggregate to improve the effectiveness of a prediction model [3] [5] [7].

Naïve Bayes: is also a simple classifier to calculate event probability. It needs a small number of data points to find relationship between the probabilities and conditional probabilities of two events. It assumes that existence of features of a class are independent of each other, and the input data is normally distributed. It is fast, scalable and effective in handling missing data [8].

Support vector machine: is another well-established and robust classification technique. Essentially it works as a hyperplane and divides the positive and negative classes in supervised learning to separate the cases of the target variables. It divides cases of two categories on two sides and tries to reach maximum margin between the two sides. If dataset contains categorical values then the use of methods such as one-hot-encoding is needed to transform them into binary values [8].

Extreme gradient boosting (XGBoost): is an improvement to already existed gradient tree boosting algorithm with increased speed and scalability than many

other ML algorithms. It is widely applied these days in ML competitions for its effective performance [3] [7].

Neural network: works like the human brain where there is a web of connected neurons. It basically comprises of number of layers; input layer, hidden layer, and output layer; with number of neurons in each layer. Each neuron in neural network algorithm takes input values and computes its weight. It then applies activation function to produce a single output value. Neural network is used to predict both continuous and categorical data. They can be effective in models where the relationship between input and output variables is non-linear [5].

2.3 Performance Analysis

The performance analysis in this paper was conducted to evaluate the overweight and obesity prediction models constructed using various machine learning algorithms in the test set using ROC-AUC, accuracy and sensitivity, and specificity. The stochastic gradient boosting machines (gbm) remained the best model to predict overweight and obesity status in the separate test data set, with ROC-AUC and accuracy values of 0.72 and 0.67, respectively. The confusion matrix was used to present the overall accuracy, sensitivity, and specificity observed in the testing set samples, which then evaluates the performance of each prediction model [3].

They evaluated model performance using K-fold cross-validation, mean average error (MAE), and Pearson's correlation coefficient (R2).The combined model showed optimal performance with the lowest mean average error (0.98, SD = 0.03) and the highest correlation (R2 = 0.72), likely owing to the greater number of patients included [4].

The paper evaluates the performance of the proposed machine learning-based models using performance metrics such as accuracy, precision, recall, and F1 score. The proposed models can predict a child's obesity category (normal, overweight, or obese) at five years of age with an accuracy of 89%, 77%, and 89%, for the three application scenarios, respectively [5].

The authors evaluated the performance of the models using various metrics such as F1, AUC, G-mean, accuracy, sensitivity, and specificity. They also trained

different models for the GWAS, EWAS, and biochemistry datasets using both the original (and imbalanced) datasets and the balanced versions of the original datasets [5].

2.4 Available Databases

The data used in this paper is genome-wide single nucleotide polymorphism (SNP) genotype and imputed data from the Framingham Heart Study (FHS) cohort. The data was downloaded from dbGaP (accession: phs000342.v18.p11) and filtered at the sample and SNP level [3].

The paper proposes three different machine learning-based techniques for predicting childhood obesity based on three different scenarios of available data. The methods are as follows [5]:

- Single well-child visit data: This method predicts obesity category (normal, overweight, or obese) at five years of age using only one well-child visit data. The proposed model uses a random forest algorithm with a 70%-30% train-test split.
- Multiple early well-child visit data: This method predicts obesity category at five years of age using multiple well-child visits under the age of two. The proposed model uses a random forest algorithm with a 70%-30% train-test split.
- Multiple random well-child visit data: This method predicts obesity category at five years of age using multiple random well-child visits under the age of five. The proposed model uses a random forest algorithm with a 70%-30% train-test split.

The paper uses multi-omics data from the PUBMEP project, a longitudinal research study that follows children with and without obesity from pre-puberty to puberty to evaluate the prevalence of metabolic syndrome and the progression of related cardiometabolic risk factors. The data includes GWAS, EWAS, clinical, anthropometric, and biochemistry data from 90 Spanish children, which were employed as predictors for the insulin resistance (IR) status at the pubertal stage [6].

The authors of the paper used saliva samples collected from children to evaluate the DNA methylation levels of the genes NRF1, FTO, and LEPR using real-time quantitative PCR-based multiplex MethyLight technology. ALU was used as a reference gene in every well to normalize the input DNA [7].

Chapter 3: Methodology and Research Plan

3.1 Data collection

Epigenetic modifications can influence gene expression without altering the underlying DNA sequence. Changes in epigenetic markers can affect the expression of genes involved in metabolism and fat storage. Therefore we choose methylation data of children for our research. To collect this data we use gene expression omnibus repository. This is public data repository and maintain many type of molecular data. It is maintained by the National Center for Biotechnology Information (NCBI).

Our data set is "Alterations of DNA Methylation Profile in Peripheral Blood of Children with Simple Obesity". This data set generated using "Bisulphite converted DNA from the 72 samples were hybridised to the Illumina Infinium MethylationEPIC BeadChip". This data set contain Genome wide DNA methylation profiling of peripheral blood samples from 41 children with simple obesity and 31 normal controls. The Illumina Infinium MethylationEPIC BeadChip (Illumina 850k, San Diego, CA) was used to obtain DNA methylation profiles across greater than 850,000 CpG sites across the genome. Samples included 31 normal and 41 obesity peripheral blood.

Each row represents a specific CpG site (identified by an ID_REF), and each column represents a different sample. The values in the table represent the DNA methylation levels at each CpG site for each sample, ranging from 0 to 1. Also column include "Detection Pval," which represents the p-value for detecting the methylation level at each CpG site. A p-value of 0 suggests that the methylation level was detected with high confidence, while a higher p-value indicates a lower confidence level.

3.2 Data preprocessing

The next phase in the process of data analysis and machine learning is data pretreatment, which is very significant. To guarantee that the data is in an appropriate format for additional analysis and modeling, it comprises a number of crucial processes [5] [7].

(1) Data cleaning: This phase entails locating and dealing with inaccurate or missing data. Different imputation methods can be used to fill in missing data, and inaccurate data can be fixed or eliminated from the dataset.

(2) Data normalization: Normalization is the process of uniformly scaling all numerical variables. This is crucial because the learning process of the model could be dominated by features of various sizes. Z-score normalization and Min-Max scaling are popular normalizing methods.

(3) Data Transformation: Occasionally, certain features might not be in the right format for analysis. To make the data better suited for modeling, data transformation techniques, including log transformations, can be used.

(4) Eliminating Duplicates: It's crucial to locate and eliminate duplicates from the dataset because they can skew results.

(5) Feature Selection: Not all dataset features may be equally useful for analysis or modeling. By recognizing and retaining only the features that are most pertinent to the work at hand, features can be selected.

(6) Handling Outliers: An analysis's or a model's performance may be greatly impacted by outliers, which are extreme values. Techniques like truncation, capping, or imputation can be used to treat them.

(7) Encoding Categorical Variables: Typically, inputs for machine learning models must be numerical. Therefore, using methods like one-hot encoding or label encoding, categorical variables are transformed into numerical representations.

(8) Handling Skewed Data: Biased models can result from skewed data, when one class or value dominates the distribution. This problem can be solved using methods like resampling (oversampling or undersampling). The accuracy and effectiveness of the final models and the insights drawn from the data are directly impacted by the quality of data preparation.

3.3 Differential methylation analysis

Finding CpG sites with noteworthy differences in DNA methylation levels between different samples or populations is an important step in epigenetic research. Differential methylation analysis procedures Choosing the CpG sites that will be examined for differential methylation is the first step. Depending on the research issue, we frequently concentrate on specific regions of interest or genome-wide CpG sites [5].

(2) Sample Grouping: The samples are divided into different groups depending on the experimental design, such as treatment vs. control, diseased vs. healthy, or different tissue types.

(3) Multiple Testing Correction: The likelihood of generating false positives increases when comparing thousands of CpG sites at once. To solve this issue, multiple testing correction methods like Bonferroni, Benjamini-Hochberg, or false discovery rate (FDR) control are used to alter the p-values for statistical significance. We frequently offer effect sizes (such as mean difference or fold change) to quantify the extent of the methylation variations between the groups in addition to statistical significance.

(4) Results Visualization and Interpretation: To emphasize significantly differentially methylated CpG sites, the results are often visualized using plots like heatmaps or volcano plots. The findings are then discussed by we in light of their biological or clinical importance.

3.4 Clustering and visualization

Clustering and visualization are essential steps in the analysis of DNA methylation data. These steps help to identify patterns or subgroups of samples or CpG sites based on their methylation profiles and provide valuable insights into the underlying biology [5] [9].

3.4.1 Clustering Samples:

Based on their methylation profiles, samples are classified together in this step. There are many clustering algorithms available, including DBSCAN (Density-

Based Spatial Clustering of Applications with Noise), k-means clustering, and hierarchical clustering. Each sample is shown as a vector of the methylation levels at particular CpG sites. The identification of samples with comparable methylation patterns—which may relate to certain biological or clinical subgroups—is made possible by clustering.

3.4.2 Clustering CpG Sites:

Similar to this, CpG sites can be grouped according to how often they are methylated in various samples. This makes it easier to find CpG sites or co-methylated regions with similar methylation activity. Understanding the regulatory areas or functional modules impacted by changes in DNA methylation can be helped by clustering CpG sites [5] [11].

3.4.3 Heatmaps:

A popular visualization method for showing clustered data is the usage of heatmaps. Heatmaps are used to show the methylation levels of samples across CpG sites or the opposite in the context of DNA methylation study. It is possible to visually recognize patterns of methylation changes across samples or CpG sites since each cell in the heatmap is color-coded according to the methylation level [2] [5].

3.4.4 Scatter Plots:

For displaying relationships between two variables, scatter plots are helpful. Scatter plots can be used in DNA methylation studies to show the methylation levels of two samples or two CpG sites. Understanding the similarities or variations in the methylation patterns between certain samples or CpG sites is made easier with the aid of this depiction [7].

3.4.5 Principal Component Analysis (PCA) and t-SNE:

High-dimensional DNA methylation data is frequently visualized in two or three dimensions using these dimensionality reduction approaches. They enable to

visualize the overall structure and patterns of variation in the data and identify potential clusters or subgroups [7].

3.5 Functional enrichment analysis

Functional enrichment analysis is a important step in understanding the biological significance of differentially methylated CpG sites. By identifying if these sites are enriched in specific genomic regions, gene pathways, or biological functions, we can gain insights into the potential functional implications of the observed methylation changes [5].

3.6 Model Development:

3.6.1 Choose Suitable Machine Learning Algorithms

Each algorithm has its strengths and weaknesses, and different algorithms may perform better on different datasets. Considering that dealing with a binary classification task (obese vs. non-obese), several machine learning algorithms are suitable for this task

- Decision Trees: A simple and interpretable algorithm that can capture non-linear relationships between features and the target variable.
- Random Forests: An ensemble learning method that combines multiple decision trees to improve performance and reduce overfitting.
- Support Vector Machines (SVM): A powerful algorithm that finds the optimal hyperplane to separate the data into different classes.
- Neural Networks: Deep learning models that can handle complex relationships and patterns in the data. They require a larger dataset and may need more computational resources.

3.6.2 Train the Model

Train the selected machine learning algorithm using the training dataset. The program will discover patterns and connections between the molecular characteristics and obesity status throughout this step.

3.6.3 Evaluate Model Performance

Once the model is trained and optimized, evaluate its performance on the testing set using appropriate evaluation metrics.

- Accuracy: The proportion of correctly classified instances over the total instances.
- Precision: The proportion of true positive predictions out of all positive predictions. It indicates the model's ability to avoid false positives.
- Recall (Sensitivity or True Positive Rate): The proportion of true positive predictions out of all actual positive instances. It represents the model's ability to detect positive cases.
- F1 Score: The harmonic mean of precision and recall. It provides a balance between precision and recall.
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): A metric that evaluates the model's ability to discriminate between classes, especially when dealing with imbalanced datasets.

3.7 Model Optimization and Testing

3.7.1 Hyperparameter Tuning

Tune the hyperparameters of machine learning algorithms (such as random forests, support vector machines, and neural networks). Hyperparameters, which are variables not learned during training but which have a big effect on the model's performance. Find the ideal configuration that maximizes performance on the validation set by experimenting with various combinations of hyperparameters using approaches like grid search or randomized search.

3.7.2 Ensemble Techniques

To enhance overall performance, think about combining the predictions of various models using ensemble approaches. Predictive accuracy can be improved by using ensemble approaches like bagging and boosting that reduce overfitting. For instance, to develop a more powerful final classifier, we can combine various models using boosting (e.g., AdaBoost) or multiple decision trees (random forests).

3.7.3 Cross-Validation

To evaluate the model's performance and guarantee its generalizability over multiple subsets of data, do cross-validation (e.g., k-fold cross-validation). As a result, overfitting is less likely to occur and a more accurate estimation of the model's actual performance is produced.

3.7.4 Regularization

To avoid overfitting in models with several features (high-dimensional data), think about applying regularization techniques like L1 or L2 regularization. Regularization alters the objective function of the model by including penalty terms, which dissuades it from leaning too heavily on any one feature.

3.7.5 Iteration and Refinement

Iterations are common in the process of model optimization. Till we obtain adequate performance, keep experimenting with various algorithms, hyperparameters, and feature sets. Keep a record of the adjustments made throughout each iteration and contrast the outcomes to find areas for improvement.

3.7.6 Testing on an Independent Dataset

Using a completely different dataset to evaluate the performance of the final optimized model. The model construction and hyperparameter tweaking phases shouldn't have used this dataset. Testing using a different dataset simulates the application of the model to fresh, unexplored data in the actual world.

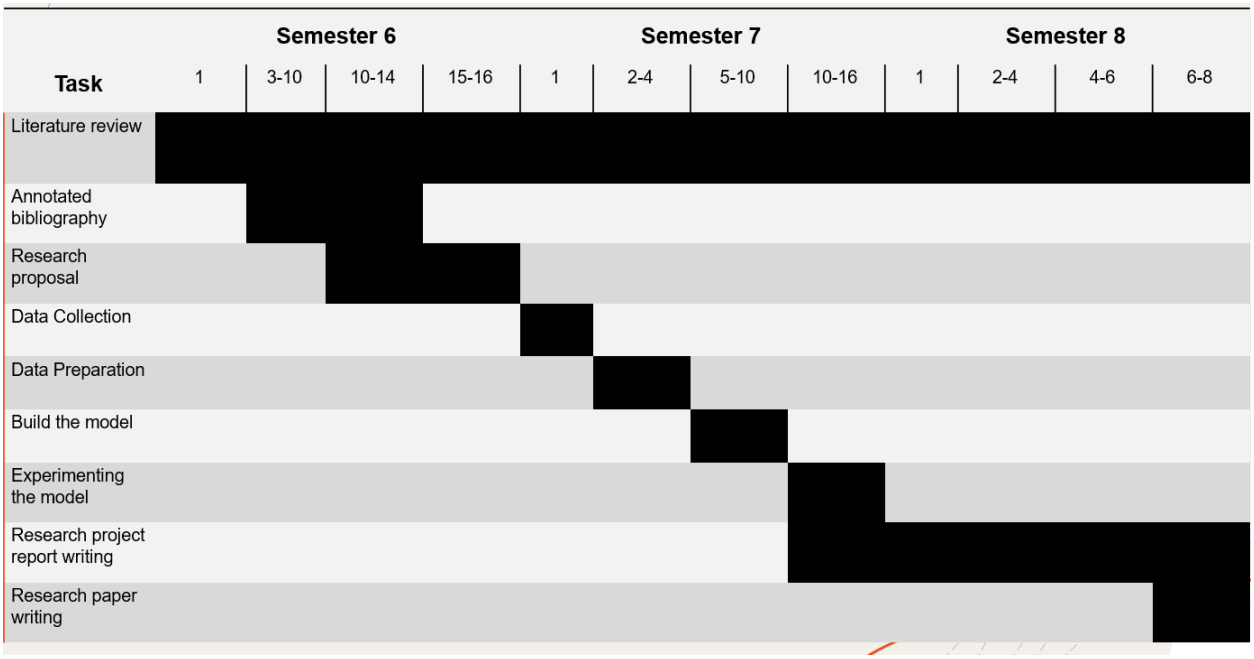
3.7.7 Evaluate Performance on the Independent Dataset

Utilize the independent dataset to apply the model, then assess its performance using the same metrics as cross-validation. This stage offers a fair evaluation of the model's adaptability to fresh data.

3.7.8 Interpretation and Reporting

Interpret the findings of testing on the independent dataset and cross-validation. Report the performance metrics, which should include both the metrics derived from the independent dataset and the average metrics from cross-validation. In addition, explore potential areas for improvement and offer insights into the model's advantages and disadvantages.

3.8 TimeLine



Chapter 4: PROGRESS TO DATE

4.1 Literature Review

With regard to our research topic, we have thoroughly studied a variety of sources, including websites, and research publications. A review of the literature will be done as the investigation progresses.

4.2 Database Collection

4.2.1 Molecular Data selection

Transcriptomic, Proteomic, Metabolomic, and other molecular data are involved in obesity, as well as genomic, epigenomic (DNA methylation, Histone Modifications), transcriptomic, and proteomic data. Since we only had access to one of these data sets, DNA methylation was the one we chose.

Bibliography

[1] H. Choquet and D. Meyre, "Molecular basis of obesity: current status and future prospects," *Current genomics,* vol. 12, p. 154–168, 2011.

[2] M. Safaei, E. A. Sundararajan, M. Driss, W. Boulila and A. Shapi, "A systematic literature review on obesity: Understanding the causes & consequences off obesity and reviewing various machine learning approcahes used to predict obesity," *Computers in biology and medicine,* vol. 136, no. no. 1, p. 104754, 2021.

[3] J. Shon and D. Pharm, "Using Machine Learning to Predict Adult Obesity Prevalence in US Counties".

[4] J. J. Milner, Z.-F. Chen, J. Grayson and S.-Y. P. K. Shiao, "Obesity-Associated Differentially Methylated Regions in Colon Cancer," *Journal of Personalized Medicine,* vol. 12, p. 660, 2022.

[5] Y.-C. Lee, J. J. Christensen, L. D. Parnell, C. E. Smith, J. Shao, N. M. McKeown, J. M. Ordovás and C.-Q. Lai, "Using machine learning to predict obesity based on genome-wide and epigenome-wide gene–gene and gene–diet interactions," *Frontiers in Genetics,* vol. 12, p. 783845, 2022.

[6] E. R. Cheng, R. Steinhardt and Z. Ben Miled, "Predicting childhood obesity using machine learning: Practical considerations," *BioMedInformatics,* vol. 2, p. 184–203, 2022.

[7] M. Alkhalaf, P. Yu, J. Shen and C. Deng, "A review of the application of machine learning in adult obesity studies," *Applied Computing and Intelligence,* vol. 2, p. 32–48, 2022.

[8] P. Patel, V. Selvaraju, J. R. Babu, X. Wang and T. Geetha, "Racial Disparities in Methylation of NRF1, FTO, and LEPR Gene in Childhood Obesity," *Genes,* vol. 13, p. 2030, 2022.

[9] P. K. Mondal, K. H. Foysal, B. A. Norman and L. S. Gittner, "Predicting Childhood Obesity Based on Single and Multiple Well-Child Visit Data Using Machine Learning Classifiers," *Sensors,* vol. 23, p. 759, 2023.

[10] Á. Torres-Martos, M. Bustos-Aibar, A. Ramírez-Mena, S. Cámara-Sánchez, A. Anguita-Ruiz, R. Alcalá, C. M. Aguilera and J. Alcalá-Fdez, "Omics Data Preprocessing for Machine Learning: A Case Study in Childhood Obesity," *Genes,* vol. 14, p. 248, 2023.