

**A COMPUTATIONAL FRAMEWORK FOR IDENTIFYING
GENE SIGNATURES USING HUMAN ADIPOSE AND
SKELETAL MUSCLE TRANSCRIPTOMES**

**A THESIS
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF**

**MASTER OF TECHNOLOGY
In
BIOINFORMATICS**

By:

**KANAN
(Enrollment No. 24/BIO/14)**

Under the Supervision of

**Prof. Asmita Das
Delhi Technological University**



**Department of Biotechnology
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042
May 2026**



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE DECLARATION

I, **KANAN**, Roll No. **2K24/BIO/014**, student of **M.Tech Bioinformatics**, hereby declare that the thesis titled "**A COMPUTATIONAL FRAMEWORK FOR IDENTIFYING GENE SIGNATURES USING HUMAN ADIPOSE AND SKELETAL MUSCLES TRANSCRIPTOMES**" submitted by me to the Department of Biotechnology, Delhi Technological University, Delhi, in partial fulfilment of the requirements for the award of the degree of **Master of Technology**, is an original research work carried out by me under proper guidance and supervision and has not been copied from any source without due acknowledgement and citation. I further declare that this work has not previously formed the basis for the award of any degree, diploma, fellowship, associateship, or any other similar academic recognition in this or any other institution or university, and I take full responsibility for the authenticity and originality of the work presented in this thesis.

PLACE: DELHI

DATE: 27/05/2026

Kanan

KANAN

This is to certify that the student has incorporated all the corrections suggested by the examiner in the thesis and the statement made by the candidate is correct to the best of our knowledge.

[Signature]
27/5/26
Signature of
Supervisor



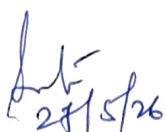
DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultpur, Main Bawana Road, Delhi-42

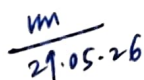
CERTIFICATE

I hereby certify that the thesis titled "**A COMPUTATIONAL FRAMEWORK FOR IDENTIFYING GENE SIGNATURES USING HUMAN ADIPOSE AND SKELETAL MUSCLES TRANSCRIPTOMES**" submitted by **KANAN**, Roll No. **2K24/BIO/14**, student of **M.Tech Bioinformatics**, Department of Biotechnology, Delhi Technological University, Delhi, in partial fulfilment of the requirements for the award of the degree of **Master of Technology**, is a Bonafide record of original research work carried out by the student under my guidance and supervision. To the best of my knowledge and belief, the work embodied in this thesis has not been submitted, either in part or in full, for the award of any degree, diploma, fellowship, associateship, or any other similar academic recognition by this or any other university or institution.


Dr. Prof. Asmita Das

Supervisor

Department of Biotechnology, Delhi
Technological University


29.05.26

Prof. Yasha Hasija

Head of Department

Department of Biotechnology
Delhi Technological University

Date: 27/05/2026



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my project guide and mentor, Prof. Asmita Das, for her invaluable guidance, constant encouragement, and continuous support throughout the course of this research work. Her insightful suggestions, scientific perspective, and dedication towards academic excellence greatly contributed to the successful completion of this project. Her mentorship not only strengthened my understanding of computational biology and antimicrobial resistance research but also motivated me to approach scientific problems with analytical thinking and confidence.

I would also like to extend my heartfelt thanks to the Ph.D. scholars in the laboratory, Ms. Monika, Mr. Yuvraj, and Ms. Simran, for their valuable guidance, technical assistance, and constant support during the project. Their constructive discussions, suggestions, and cooperative nature significantly enhanced my learning experience and helped me overcome various challenges encountered during the research work. I am equally grateful to the faculty members of the Department of Biotechnology, Delhi Technological University, whose teachings and academic support throughout the M.Tech programme helped me build a strong foundation in bioinformatics and computational research.

Finally, I would like to express my deepest gratitude to my family, friends, and batchmates for their unwavering encouragement, emotional support, and belief in my abilities throughout this academic journey. Their continuous motivation and support have always been a source of strength and inspiration, helping me successfully complete this work and progress confidently in my academic and professional pursuits.

ABSTRACT

Obesity is a multi-cellular metabolic syndrome characterised by dysregulated adipose tissue growth, mitochondrial bioenergetics and disrupted glucose and lipid homeostasis. Although transcriptome-scale profiling provides a thorough understanding of these changes, the accuracy of conventional statistical methods is limited by the nonlinear and high-dimensional structure of gene expression data. To address these challenges, an integrated machine-learning and explainable artificial intelligence framework was developed to discover transcriptomic predictors of obesity in human skeletal-muscle and adipose tissue. Ensemble classification models i.e. Random Forest and Extreme Gradient Boosting were applied to normalised gene expression matrices derived from publicly available datasets(GSE22309 and GSE103766) with systematic hyperparameter optimization. SHapley Additive exPlanations(SHAP) promote model interpretability, enabling quantitative measure of each gene contributions to classification outcomes. Key transcripts such as BDH1, MRPS31, RARS2, STOML2, AKAP10 and UCP3 were aligned with mitochondrial metabolic pathways involving ketone oxidation, ribosomes function, cristae stability and proton leak modulation. Network analysis identified NUCB2, SRSF1 and ACTG2 as the key topological nodes while Enrichment analysis using Enricher revealed significant alterations in oxidative phosphorylation interactions, reactive oxygen and nitrogen mediated signalling pathways. Overall this study shows the significance of an xAI enabled framework for identifying biologically significant transcriptomic signatures related to obesity.

TABLE OF CONTENTS

Title	Page no.
Candidate's Declaration.....	(i)
Certificate.....	(ii)
Acknowledgement.....	(iii)
Abstract.....	(iv)
Contents.....	(v-vi)
List of Figures	(vii)
List of Tables.....	(vii)
List of Abbreviations	(ix)
 CHAPTER – 1	
INTRODUCTION	1-17
1.1. Background	1
1.2. Literature Review	5
1.2.1 Molecular Basis and Pathophysiology of Obesity.....	5
1.2.2 Adipose Tissue Dysfunction in Obesity.....	5
1.2.3 Skeletal Muscle Metabolism and Insulin Resistance in Obesity.....	6
1.2.4 Mitochondrial Dysfunction and Oxidative stress in Obesity.....	8
1.2.5 Transcriptomics and Biomarker Discovery in Obesity.....	9
1.2.6 ML Approaches in Obesity and Transcriptomic Research.....	10
1.2.7 XAI and SHAP in Biomarker Discovery.....	12
1.2.8 PPI Networks and Pathway Enrichment Analysis.....	13
1.2.9 Research Gap.....	16
1.3 Objective.....	17
 CHAPTER – 2	
MATERIALS AND METHODOLOGY	18-31
2.1. Dataset Acquisition.....	18
2.2. Data Preprocessing and Transcriptomic Integration.....	20

2.3. Construction of Feature Matrix and Target Variable.....	22
2.4. Machine Learning Model Development.....	23
2.5. Model Evaluation and Performance Assessment	26
2.6. XAI Analysis with SHAP	27
2.7. Protein Protein Interaction Network Analysis.....	29
2.8. Functional Pathway Enrichment Analysis.....	31

CHAPTER – 3

RESULTS 33-47

3.1. Implementation of Machine Learning Models.....	33
3.2. Analysis of the Confusion Matrix of Random Forest Classifier.....	34
3.3. SHAP Based Global Feature Importance Analysis.....	37
3.4. Cohort-Level SHAP Interpretation of Transcriptomic Features	39
3.5. Local Interpretability of Obesity Classification Using SHAP	41
3.6. Protein–Protein Interaction Network Analysis of Identified Biomarkers.....	43
3.7. Functional Pathway Enrichment Analysis of Transcriptomic Biomarkers.....	45
3.8. Integrated interpretation of transcriptomic findings	47

CHAPTER - 4

DISCUSSION 49-51

4.1. Biological Interpretation of Obesity-Associated Transcriptomic Signatures.....	49
4.2. Comparison with Previous Obesity Transcriptomic Studies.....	50
4.3. Biological implications of mitochondrial dysfunction in obesity.....	51

CHAPTER - 5

CONCLUSION AND FUTURE SCOPE.....52

REFERENCES.....56

PLAGARISM REPORT.....65

LIST OF FIGURES

Figure No.	Figure Caption	Page Number
Fig. 1	Workflow of the study	19
Fig. 2	Acquisition and Preprocessing of Gene Expression Data	21
Fig. 3	Machine Learning Pipeline for Classification	25
Fig. 4	Post Modelling Analysis and Biological Interpretation	27
Fig. 5	Random Forest confusion matrix showing robust classification performance between control and obesity samples	36
Fig. 6	SHAP global feature importance bar plot highlighting the top 20 obesity-associated transcriptomic biomarkers	38
Fig. 7	SHAP beeswarm plot illustrating cohort-level distribution and directional contribution of top-ranked transcriptomic features	40
Fig.8	SHAP waterfall plot showing local feature contributions driving an individual obesity prediction	42
Fig.9	STRING–Cytoscape protein–protein interaction network identifying NUCB2, SRSF1, and ACTG2 as central hub genes	44
Fig.10	Enrichr KEGG/Reactome pathway enrichment analysis revealing mitochondrial dysfunction and metabolic dysregulation pathways in obesity	46

LIST OF TABLES

TABLE NO.	TITLE OF TABLES	PAGE NO.
Table 1	Classification Performance of Machine Learning Models	34
Table 2	Confusion Matrix Interpretation of Random Forest Model	35
Table 3	Major Enriched Biological Pathways Identified in Obesity	47

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
AKAP10	A-Kinase Anchoring Protein 10
ATP	Adenosine Triphosphate
ACTG2	Actin Gamma 2
AUC	Area Under the Curve
XGBoost	Extreme Gradient Boosting
SHAP	SHapley Additive exPlanations
BDNF	Brain-Derived Neurotrophic Factor
BMI	Body Mass Index
ML	Machine Learning
cDNA	Complementary Deoxyribonucleic Acid
SVM	Support Vector Machine
CEL	Cell Intensity File
Cytoscape	Cytoscape Software for Network Analysis
KNN	K-Nearest Neighbour
KEGG	Kyoto Encyclopedia of Genes and Genomes
JNK	c-Jun N-terminal Kinase
NCBI	National Center for Biotechnology Information
GSE	GEO Series
GEO	Gene Expression Omnibus
PCA	Principal Component Analysis
NUCB2	Nucleobindin 2
NF- κ B	Nuclear Factor Kappa B
LR	Logistic Regression
DT	Decision Tree
ROS	Reactive Oxygen Species
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins

CHAPTER 1

INTRODUCTION

1.1 Background

Obesity is a chronic multifactorial metabolic condition with an excess fat accumulation that negatively influences the functioning of the body and promotes the development of various metabolic and cardiovascular diseases [1,2]. It is caused by a long-term positive energy balance, when energy intake is more than energy expenditure over time, but has a much more complicated pathogenesis, which is dependent on interaction between genetic, environmental, behavioural and molecular factors [3,4]. In recent decades, obesity has become one of the greatest public health problems worldwide due to its increased prevalence in both industrialized and developing countries [5]. This has led to a growing prevalence of obesity-related conditions such as type 2 diabetes mellitus, cardiovascular disease, hypertension, dyslipidaemia, metabolic syndrome and some types of cancer [6,7]. BMI is widely used to diagnose obesity in adults ($\text{BMI} > 30 \text{ kg/m}^2$) [5]. BMI is an easy to measure epidemiological indicator but also does not reflect the distribution of body fat or metabolic differences between individuals [8]. Increased adiposity of various body compartments, especially abdominal fat, has been reported to have a stronger relationship with metabolic dysfunction and insulin resistance than with total body weight [9]. Therefore obesity is no longer viewed as simply over accumulation of body fat, but is now considered to be a systemic metabolic disease with a large amount of physiological and molecular abnormalities [10]. The biological mechanisms underlying obesity involve dysregulation of energy homeostasis controlled by neuroendocrine, metabolic and molecular signaling networks [11]. The hypothalamus, gastrointestinal tract, adipose tissue and endocrine

signaling pathways are involved in the regulation of appetite and energy balance [12]. Leptin, ghrelin and insulin are hormones that are important for hunger, satiety and glucose metabolism [13]. Adipose tissue produces a hormone called leptin, which under normal physiological conditions signals fullness to the hypothalamus which in turn inhibits food intake [14]. In most cases, however, increased levels of leptin do not seem to provide sufficient satiety to curb appetite and lead to ongoing energy intake and weight gain (leptin resistance) [15]. Although adipose tissue was once seen as a passive lipid storage organ, modern research has identified it as a dynamic endocrine organ that plays a role in metabolic control [16]. The biologically active molecules produced by adipose tissue are called adipokines, which include leptin, adiponectin, inflammatory cytokines and resistin, all of which are responsible for the regulation of insulin sensitivity, lipid metabolism, and systemic inflammation [17]. Obesity is associated with elevated levels of pro-inflammatory adipokines and decreased anti-inflammatory and insulin-sensitizing adipokines including adiponectin, which are responsible for the changes in adipokine secretion profiles in obesity [18]. These molecular alterations play an important role in the processes of metabolic dysfunction and the development of insulin resistance [19].

Another important metabolic tissue that is also tightly coupled to obesity-associated pathophysiology is skeletal muscle, which in humans represents around 70-80% of insulin mediated glucose uptake. Glucose and fatty acids are effectively used by skeletal muscle under physiological conditions for energy homeostasis [21]. Obesity however, has a negative effect on insulin signalling pathways in skeletal muscle, resulting in decreased glucose uptake and mitochondrial oxidative efficiency [22]. These deficits lead to metabolic inflexibility, build up of intracellular lipid metabolites and are linked to insulin resistance and type 2 diabetes mellitus [23]. Mitochondrial dysfunction has been well recognized in metabolic abnormalities of obesity

[24]. The mitochondria are essential in the process of oxidative phosphorylation and to control the homeostasis of reactive oxygen species (ROS) [25]. Mitochondrial dysfunctions in obese subjects include reduced electron transport chain (ETC) activity, diminished ATP production and increased production of reactive oxygen species (ROS) [26]. The resulting elevated oxidative stress then triggers inflammatory pathways and leads to further damage, poor insulin signaling and metabolic disorders [27]. Such abnormalities at the molecular level have been found in both adipose tissue and skeletal muscle, suggesting a systemic manifestation of metabolic dysfunction associated with obesity [28].

As adipose tissue expands it is infiltrated by immune cells, including macrophages, that release inflammatory cytokines including tumour necrosis factor alpha (TNF- α), interleukin-6 (IL-6) and interleukin-1 beta (IL-1 β), [30]. These cytokines alter the insulin receptor signaling pathways and worsen insulin resistance by triggering stress-related signaling pathways [31]. There is also an association with endothelial dysfunction, metabolic syndrome and elevated cardiovascular risk [32] that persists because of the inflammatory signalling. With the rapid development of high throughput omics technologies, understanding of molecular mechanisms involved in obesity has improved greatly [33]. In particular, the techniques of transcriptomics have allowed genome-scale analysis of changes in gene expression that are linked to obesity and metabolic disorders [34]. Large scale gene expression repositories like the Gene Expression Omnibus (GEO) offer a collection of data that can be used to systematically study molecular signatures associated with obesity [35]. Changes in genes relevant to oxidative phosphorylation, glucose metabolism, inflammatory signalling and mitochondrial homeostasis have been observed in metabolically relevant tissues, such as adipose tissue and skeletal muscle, using transcriptomic analysis [36]. The advantage of cross-tissue transcriptomic analysis is that it

provides a better understanding of the mechanisms underlying systemic diseases since the obesity condition impacts multiple metabolically connected tissues [37]. Although a lot of progress has been made in the transcriptomic research, the identification of well-established biomarker associated with obesity is still challenging because of the high dimensionality, heterogeneity and nonlinearity of biological data [38]. Complex gene–gene interaction and tissue-specific molecular relationships are not captured by conventional statistical methods [39]. Machine learning methods have thus become highly effective tools to address high-dimensional omics data analysis and to discover predictive molecular signatures [40]. Random Forest, Support Vector Machine (SVM), Gradient Boosting and Extreme Gradient Boosting (XGBoost) have been used in disease classification and biomarker identification with metabolic disorders with promising results [41]. Machine learning methods have impressive predictive power but suffer from low interpretability, which limits their use in the biological and clinical contexts [42]. Predictive models have therefore attracted significant interest in methods of making AI models explainable (XAI) [43]. Cooperative game theory-based explainability framework, SHapley Additive exPlanations (SHAP) [44] allows quantitative interpretations of feature importance and individual prediction results. These help identify biologically relevant biomarkers and boost confidence in machine learning predictions [45]. Besides the predictive modelling, another contribution of system biology to the understanding of mechanistic changes in obesity-related changes in molecular pathways has been made [46]. Biological validation of candidate biomarkers can be performed via the identification of functionally interacting proteins and dysregulated signaling pathways by using protein–protein interaction (PPI) networks and pathway enrichment analyses [47]. The above mentioned databases and tools such as STRING, Cytoscape, KEGG, Reactome and Enrichr have therefore become essential tools for interpretation of transcriptomic data in relation to metabolic diseases [48]. Ensemble machine

learning models such as Random Forest and XGBoost are used to classify obesity-associated transcriptomic profiles and SHAP-based explainability is used to identify biologically important genes that lead to the prediction outcome [44].

1.2 Literature Review

1.2.1 Molecular Basis and Pathophysiology of Obesity

Obesity is a very complex metabolic disease that is characterized by abnormalities in the regulation of many metabolic pathways, such as energy metabolism, endocrine signaling, mitochondrial homeostasis, inflammatory pathways, etc. [10,24]. The disease is a result of the combination of genetic susceptibility, environmental exposure, eating habits, sedentary lifestyle and molecular changes in the regulation of systemic metabolism [3,11]. While excess calories still play a key role, there is significant evidence that molecular and transcriptomic changes in metabolically active tissues strongly influence the progression of obesity [34]. In addition, over-expanded adipose tissue induces hypoxia, inflammatory activation and cellular stress, all of which contribute to worsen metabolic imbalance [29]. Obesity-related inflammation alters the insulin receptor signaling pathways by serine phosphorylation of insulin receptor substrate proteins, which leads to impaired glucose metabolism and hyperglycemia [31].

1.2.2 Adipose Tissue Dysfunction in Obesity

Adipose tissue is important for systemic energy homeostasis in regulating lipid storage, endocrine signaling and metabolic regulation [16,17]. Obesity-induced adipose tissue expansion leads to significant molecular and physiological changes and results in adipose tissue becoming a metabolically protective organ to a dysfunctional inflammatory environment [18].

But in obesity excess energy balance is maintained and causes excessive adipocyte expansion, causing adipocyte hypertrophy and hyperplasia [18].

Another important adipokine in regulating metabolism and insulin sensitivity is adiponectin [51]. In contrast, the blood levels of adiponectin are markedly decreased in obese people [52].

The expansion of adipose tissue is often greater than the capacity of the vessels, resulting in areas of hypoxia that trigger inflammatory signaling pathways [54]. As a result, the fat tissue hypertrophied becomes infiltrated with macrophages, neutrophils and T lymphocytes that together exaggerate inflammatory processes [30].

In obese adipose tissue, macrophages are the most prominent inflammatory immune cell [30]. Alternatively activated (M2) macrophages are known to promote anti-inflammatory functions and tissue homeostasis under lean physiological conditions [56].

Excessive collagen deposition decreases tissue compliance and adipocyte distension, and thus lipid storage capacity [59]. The fibrotic adipose tissue therefore encourages ectopic fat deposition in metabolically active tissues and helps to develop systemic insulin resistance [60].

Visceral fat has been found to be more active from an endocrine and inflammatory standpoint than other fat depots [61]. Visceral adipose tissue, due to its anatomical location, directly affects the metabolic function of the liver by secreting inflammation mediators and free fatty acid, which is why it affects liver metabolism [62].

1.2.3 Mitochondrial Dysfunction and Oxidative Stress in Obesity

Mitochondria are vital cellular organelles that play a key role in energy production, control of oxidative phosphorylation, fatty acid oxidation and maintenance of metabolic homeostasis [24,25]. As obesity is defined as a chronic state of over-consumption of nutrients and altered energy metabolism, mitochondrial dysfunction has become one of the major molecular

mechanisms of obesity-associated metabolic disorders [24]. The ability to use glucose and fatty acids in an appropriate manner to meet cellular energy requirements is related to efficient mitochondrial respiration [22]. Obesity disturbs mitochondrial bioenergetics in a way that involves a decreased activity of the electron transport chain, a decreased oxidative capacity, and impaired substrate oxidation [26]. There have been several reports of decreased mitochondrial density and diminished mitochondrial biogenesis in obese people [64]. Thus, defects in FAMO lead to increased lipid storage in the cell and hence insulin resistance [23]. Impaired mitochondrial function is well linked to the production of ROS [26]. In normal physiological conditions, naturally occurring ROS are found in mitochondria and are involved in redox homeostasis and cell signaling in a controlled manner [67]. But when the electron transport chain is overloaded by excess mitochondria, due to obesity, the amount of ROS produced is over the limit [68]. These pathways are activated and lead to impaired insulin receptor substrate signaling and have a significant role in metabolic dysfunction [57]. Mitochondrial respiration and the activity of oxidative enzymes are often impaired in the skeletal muscle tissues of obese people [72]. Changes in genes related to oxidative phosphorylation and ATP synthesis and to the organization of the mitochondrial membrane in obesity have been identified by transcriptomic studies [36]. Mitochondrial metabolic genes are decreased in skeletal muscle which reduce the ability to use glucose and fatty acids leading to reduced metabolic flexibility and systemic insulin resistance [22,72]. Likewise, during the onset of obesity, there are significant abnormalities in the mitochondria in adipose tissue [18].

Impaired mitochondria function has also been associated with changes in adipokines and inflammatory mediators secretion, which worsens obesity-related endocrine disturbances [17,18]. Some mitochondrial-associated genes have emerged as being involved in the pathogenesis of obesity, based on transcriptomic data [36]. In obesity, the genes that regulate

oxidative phosphorylation, mitochondrial ribosomal function, and fatty acid oxidation and proton leak regulation are often dysregulated in obese tissues [73]. Thus, the study of mitochondrial metabolism has become an important therapeutic avenue to obesity management and intervention in metabolic diseases [74]. Transcriptomic technologies offer genome-wide analysis of coordinated molecular changes associated with disease progression as opposed to conventional molecular technologies that target individual genes [33]. There are extreme changes of gene expression in the various metabolically active tissues of obese animals, including adipose tissue, skeletal muscle, liver and pancreatic tissue [17,20]. These tissues are involved in glucose metabolism, lipid homeostasis and endocrine signaling, and as such, transcriptomic studies offer insightful mechanistic information regarding metabolic dysfunction in obesity [36]. A large number of transcriptomic technologies have been developed to study disease-related molecular changes in a systematic manner, including high throughput microarray analysis and the RNA sequencing (RNA-seq) approach. Transcript levels across biological samples can be quantified in thousands of genes simultaneously using microarray based studies [75]. The improvements in transcriptomic resolution of RNA sequencing technologies further enhance the capability to identify changes in gene expression in an unbiased way, as well as transcript variants and novel molecular signatures [76].

1.2.4 Mitochondrial Dysfunction and Oxidative Stress in Obesity

Mitochondria are vital cellular organelles that play a key role in energy production, control of oxidative phosphorylation, fatty acid oxidation and maintenance of metabolic homeostasis [24,25]. There have been several reports of decreased mitochondrial density and diminished mitochondrial biogenesis in obese people [64]. Transcriptional coactivators like peroxisome proliferator-activated receptor gamma coactivator-1 alpha (PGC-1 α) control the expression of

genes involved in oxidative metabolism and mitochondrial respiration, and are the main regulators of mitochondrial biogenesis [65]. PGC-1 α and mitochondrial transcription factors are down-regulated in obesity, leading to a decrease in the number of mitochondria and a decrease in the efficiency of oxidative phosphorylation [66]. In normal physiological conditions, naturally occurring ROS are found in mitochondria and are involved in redox homeostasis and cell signaling in a controlled manner [67]. But when the electron transport chain is overloaded by excess mitochondria, due to obesity, the amount of ROS produced is over the limit [68]. High ROS levels result in damage to proteins, lipids, and nucleic acids and also affect mitochondrial integrity and function [69]. Oxidative stress is caused by an imbalance between the production of ROS and the antioxidant capacity, causing damage to cellular structure or dysfunction in metabolism [70]. In several tissues, obesity has always been linked to the increase of oxidative stress markers and to the decrease of antioxidant capacity [71]. Mitochondrial respiration and the activity of oxidative enzymes are often impaired in the skeletal muscle tissues of obese people [72].

1.2.5 Transcriptomics and Biomarker Discovery in Obesity

Transcriptomic obesity studies have benefited from the use of publicly available repositories [35]. The existence of transcriptomic datasets related to obesity makes it possible to carry out integrative analysis and to obtain reproducible obesity-related biomarkers in different cohorts [77]. Given the data set used for obesity, the adipose tissue and skeletal muscle sets are important since these tissues are central to energy metabolism and insulin sensitivity [18,20]. Transcriptomic analysis of these skeletal muscle tissues has shown differences in the expression of genes that control mitochondrial oxidative metabolism, glucose transport and ATP production in obese people [22,36]. Integrative transcriptomics studies give a biological

overview superior to single tissue studies as obesity impacts a multitude of interconnected metabolic tissues [38]. The analysis of the simultaneous adipose and skeletal muscle data allows the identification of conserved molecular pathways involved in metabolic dysfunction, inflammation and mitochondrial impairment [28]. Although the use of transcriptomics to discover biomarkers has made significant progress, it is difficult to discover them because of its high dimensionality and heterogeneity of gene expression data [38]. Often biological variation, technical noise and non-linear interactions between genes make the identification of candidate biomarkers using traditional statistical approaches less reproducible [39]. As a result, biomarker identification in transcriptomics has been increasingly assisted by advanced computational and machine learning methods to enhance the accuracy of biomarker identification [40].

1.2.6 Machine Learning Approaches in Obesity and Transcriptomic Research

However, conventional statistical techniques, which are used mainly to detect linear relationships, can be limited in their ability to provide insights into these types of datasets and may not be able to detect complex molecular interactions [39]. As a result, machine learning techniques have become important computational tools that allow the extraction of biologically relevant information from omics-scale data, and the discovery of molecular signatures that are predictive of complex diseases, such as obesity [40].

Supervised learning models are commonly applied in the field of obesity research for classification of obese and non-obese samples, for prediction of disease susceptibility and for the discovery of molecular biomarkers linked to metabolic dysregulation [40]. Unlike supervised learning, the unsupervised learning methods discover hidden patterns or subgroups of molecules without any specific labels and are especially suited for clustering transcriptomic profiles and detecting the heterogeneity of diseases [82].

Transcriptomic obesity studies have shown that several supervised machine learning techniques have shown to perform well [41]. Logistic Regression is among the simpler classification methods and models used to describe relationship between predictor variables, and categorical outcomes [83]. Logistic Regression, interpretable and computationally efficient, has been found to have an inferior performance in high dimensional and highly non-linear transcriptomic datasets [39]. But the performance of SVM can be degraded in very heterogenous biological data sets when the number of features becomes too high [85].

The interpretability and capability of modelling nonlinear interactions have made Decision Tree based algorithms extremely important in the field of transcriptomic biomarker discovery [86]. Decision Trees are a classification model that splits the observations sequentially along features and hence offer a clear prediction path [87]. Single decision trees can suffer from overfitting and loss of predictive stability, however [88]. To address these drawbacks, ensemble learning methods were later proposed to combine the predictions of the aggregates of multiple models [89].

Moreover, Random Forest is very useful to deal with high-dimensional data, such as transcriptomic data including thousands of variables related to the expression of genes [92].

Another highly popular ensemble machine learning method is Extreme Gradient Boosting (XGBoost) which is often used in the study of omics [93]. The prediction performance of XGBoost models is incrementally enhanced by using gradient boosting, which sequentially uses models to reduce the error of the previous models [94]. There are several transcriptomic studies that have successfully applied XGBoost to find gene signatures associated with disease and enhance the discovery of biomarkers [96].

While machine learning methods overcome these disadvantages, they can identify informative subsets of genes that are linked to disease phenotypes and exclude redundant features [97]. Thus,

computational approaches that integrate transcriptomics and machine learning have grown in use to discover obesity-associated biomarkers and therapeutic targets [98].

Though machine learning models are often highly predictive, they are often “black-box” systems that are not very interpretable [42]. Thus, an important challenge in computational biology and precision medicine is to make the predictions of machine learning more interpretable [44].

1.2.7 Explainable Artificial Intelligence (XAI) and SHAP in Biomarker Discovery

Explainable Artificial Intelligence (XAI) denotes a set of computational techniques that have been created to make ML predictions more interpretable and transparent [43]. In contrast to traditional black-box algorithms, XAI methods enable researchers to gain insights into the contribution of specific features to prediction results [44]. Interpretability is a crucial property in biomedical research, as beyond clinical trust, there is a need for biologically meaningful explanations for therapeutic decisions and mechanistic understanding.

SHapley Additive exPlanations (SHAP) [44] is one of the most commonly used explainability frameworks in transcriptomic and biomedical research.

To quantify the average contribution of each feature across all the predictions, the global SHAP analysis can be applied and it can help identify biologically important genes associated with disease phenotypes [102]. Combinations of genes explain the prediction outcomes at the individual sample level with the help of local interpretability approaches, such as SHAP force plots and waterfall plots [103].

Ensemble machine learning algorithms, including Random Forest and XGBoost, have been integrated into SHAP, greatly enhancing the biological interpretability in omics research [96].

SHAP-based analyses could be used to prioritize molecular biomarkers and to also provide a

biological explanation for computational predictions [44]. Such explainability frameworks are thus a step toward connecting predictive performance with mechanistic understanding, further improving the translatability of machine learning-based biomarker discovery [45].

Although machine learning has great potential in obesity-associated transcriptomic studies, explainable machine learning methods are relatively underutilized [98]. The majority of the previous studies have been conducted on predictive performance alone with limited biological validation of the biomarkers identified [39]. Thus, the potential for incorporating explainability within transcriptomic machine learning applications is important for enhancing the reliability, reproducibility and mechanistic understanding of molecular signatures associated with obesity [44,98].

1.2.8 Protein–Protein Interaction Networks and Pathway Enrichment Analysis in Obesity Research

These methods help in understanding the mechanisms by mapping molecular interactions, pathways of regulation, and the biological processes leading to the disease process.

Proteins do not normally act alone in biological systems, but rather are involved in highly coordinated networks of molecular interactions that control cellular homeostasis [104]. Protein–protein interaction networks provide a systematic representation of functional interactions among proteins and can identify central regulatory molecules that are associated with pathways of disease [105]. PPI network analysis has been widely used in the study of obesity to discover the genes that are hub genes in inflammation, insulin signalling and lipid metabolism and mitochondrial function [106].

A hub gene is a highly interconnected molecular hub that often has significant regulatory effects on biological pathways [107]. The genetic signature of obesity has been determined by several

obesity-related transcriptomic studies, which have been able to identify central hub genes that are related to mitochondrial respiration, oxidative phosphorylation and inflammatory signaling [108]. The presence of hub genes means that they control multiple downstream molecular mechanisms and may be suitable therapeutic targets and clinically relevant biomarkers [109].

STRING combines experimental data, computational predictions and literature-based evidence to construct detailed molecular interaction networks [111]. Machine learning or differential expression analysis can then be applied to identify candidate genes from transcriptomes that can be mapped in STRING to discover functionally connected biological pathways [112].

After network generation, Cytoscape is often used to visualize and analyze protein interaction networks [113]. These analyses help to prioritize biologically important genes that could be related to the pathogenesis of obesity [115].

The biological interpretation of transcriptomic data can be further improved using functional enrichment analysis, which can elucidate the pathways that are strongly correlated with candidate genes [47]. Transcriptomic analyses frequently lead to the identification of gene sets that are known or suspected to be involved in pathways of metabolism, oxidative stress, mitochondrial respiration, inflammatory signaling and glucose homeostasis [36]. Functional enrichment, in turn, can give mechanistic understanding of disease progression and biological validation of computational predictions [46].

There are multiple pathway databases available for enrichment analysis of obesity studies. Enrichment analyses based on KEGG databases have been consistently found to include insulin resistance, oxidative phosphorylation, fatty acid metabolism and inflammatory signaling pathways as obesity-related pathways [117]. Likewise, detailed pathway annotations on molecular interactions, cellular processes and metabolic pathways are provided by the Reactome database [118].

Another popular functional enrichment platform is Enrichr, which combines multiple pathway databases into a single platform for convenient biological interpretation of transcriptomic data [119]. Enrichr is able to identify pathways that are overrepresented in the set of candidate genes and statistically significant using enrichment scoring and statistical significance testing [120]. The pathways associated with mitochondrial dysfunction, reactive oxygen species signaling, cytokine-mediated inflammation and glucose metabolism are frequently reported to be enriched in transcriptomic studies of obesity [36,121].

Most importantly, systems biology validation enhances the reliability of machine learning-based framework for biomarker discovery [46]. Occasionally, features that are statistically significant but of no biological relevance can be identified by computational predictions only [42]. Combining PPI network analysis and pathway enrichment thus increases the confidence of discovered biomarkers that they play a role in biologically relevant molecular pathways [47].

Although numerous molecular pathways have been explored in obesity, there are a number of critical considerations that are not addressed in existing obesity biomarker discovery strategies [39]. The molecular interactions within systemic networks that play a role in obesity pathogenesis have been largely neglected in most obesity-related transcriptomic studies, which have been studying adipose tissue or skeletal muscle alone so far [37]. Because obesity is a multisystem metabolic disease that impacts the interconnected tissues, studies involving a single tissue may not capture the conserved transcriptomic signatures associated with metabolic dysfunction in the system [122].

Current transcriptomic studies have mainly focused on using differential gene expression analysis and classical statistical methods to identify biomarkers [38]. These approaches yield meaningful biological information, but they do not well represent the nonlinear molecular interactions and complex gene–gene interactions found in high-dimensional biological data sets

[39]. This means that the reproducibility and predictive power of the biomarkers identified are often limited [40].

The challenges this poses have been met with machine learning approaches, which are able to describe complex nonlinear relationships and process multidimensional transcriptomic datasets [41]. In previous machine learning studies on obesity, however, most studies have focused on prediction accuracy at the expense of limited biological interpretability [42]. Several machine learning algorithms, however, are “black box” in nature, which reduce biological confidence and clinical applicability of findings from computing [43].

Additionally, there are only some limited studies that systematically applied the EAI concepts, such as SHAP, to the field of obesity transcriptomic studies [44]. Current computational programs do not usually give a detailed description of the mechanisms through which the individual genes give rise to the classification outcome for obesity [98].

1.2.9 Research Gap

Also, biological validation studies at the systems level, by combining machine learning predictions with protein interaction networks and pathway analysis, have not been extensively used in obesity transcriptomic studies [46]. Numerous studies report candidate biomarkers without adequate validation of their involvement in functionally relevant metabolic pathways [47]. This limitation can hinder the ability to understand the mechanism and can limit the translational value of identified biomarkers [109].

Thus, there is still a significant demand for the development of an integrated computational framework that integrates multi-tissue transcriptomic analysis, ensemble machine learning, explainable artificial intelligence and systems biology validation to identify biologically interpretable obesity-associated molecular signatures [37,44].

1.3 Objectives

The objectives of the present study were to:

1. To acquire and preprocess publicly available transcriptomic data from skeletal muscle and adipose tissue of obese individuals GSE22309 and GSE103766 [35].
2. To apply SHAP to explainable AI in the identification of obesity classification outcomes with significant genes for application in a biological context [44].
3. To discover biologically meaningful transcriptomic biomarkers and molecular pathways involved in the progression of obesity and metabolic dysfunction [36].

CHAPTER 2

MATERIALS AND METHODOLOGY

2.1 Dataset Acquisition

The selection of two obesity-associated transcriptomic datasets (GSE22309 and GSE103766) for the present investigation was based on their biological relevance to metabolic dysfunction associated with obesity and tissue specific importance in energy homeostasis [35,37]. These datasets capture the metabolically active tissues that are closely linked to glucose metabolism, insulin sensitivity, and systemic metabolism, thus providing a suitable context for integrated obesity transcriptomic analysis [20,37].

The GSE22309 data contained human skeletal muscle transcriptomic profiles [20]. Many molecular changes associated with obesity in skeletal muscle have been strongly correlated with mitochondrial dysfunction, decreased oxidative metabolism and insulin resistance [22,72]. Thus, transcriptomic analysis of skeletal muscle enabled the identification of molecular signatures associated with obesity that are related to reduced glucose utilization and impaired mitochondrial metabolic function [36].

The GSE103766 dataset was derived from transcriptomics data of human adipose tissue samples [17]. Adipose tissue acts as an endocrine and metabolic organ of lipid storage, adipokine signaling and systemic inflammatory response [16]. Obesity-induced adipose tissue remodeling has been correlated with increased inflammation, altered expression of adipokines and mitochondrial dysfunction, which all play a significant role in the development of insulin resistance and in the progression of metabolic syndrome [18,29].

For an extensive understanding of the molecular mechanisms involved in obesity, beyond tissue-specific biological changes, selection of cross-tissue transcriptomic datasets was performed [37].

The datasets were carefully chosen for the following inclusion criteria: (i) obesity and control samples were available for both (ii) relevance to metabolically active human tissues, (iii) publicly accessible normalized gene expression matrix and metadata, and (iv) adequate sample representation for computational modeling and subsequent transcriptomic analysis [35,38]. Use of publicly available datasets further enhanced reproducibility, transparency and cost-effectiveness of the present computational framework [35].

After the selection of the datasets, raw transcriptomic matrices and platform annotation files were downloaded from GEO for further pre-processing, normalization and analysis for the identification of biomarkers using machine learning tools [35,40].

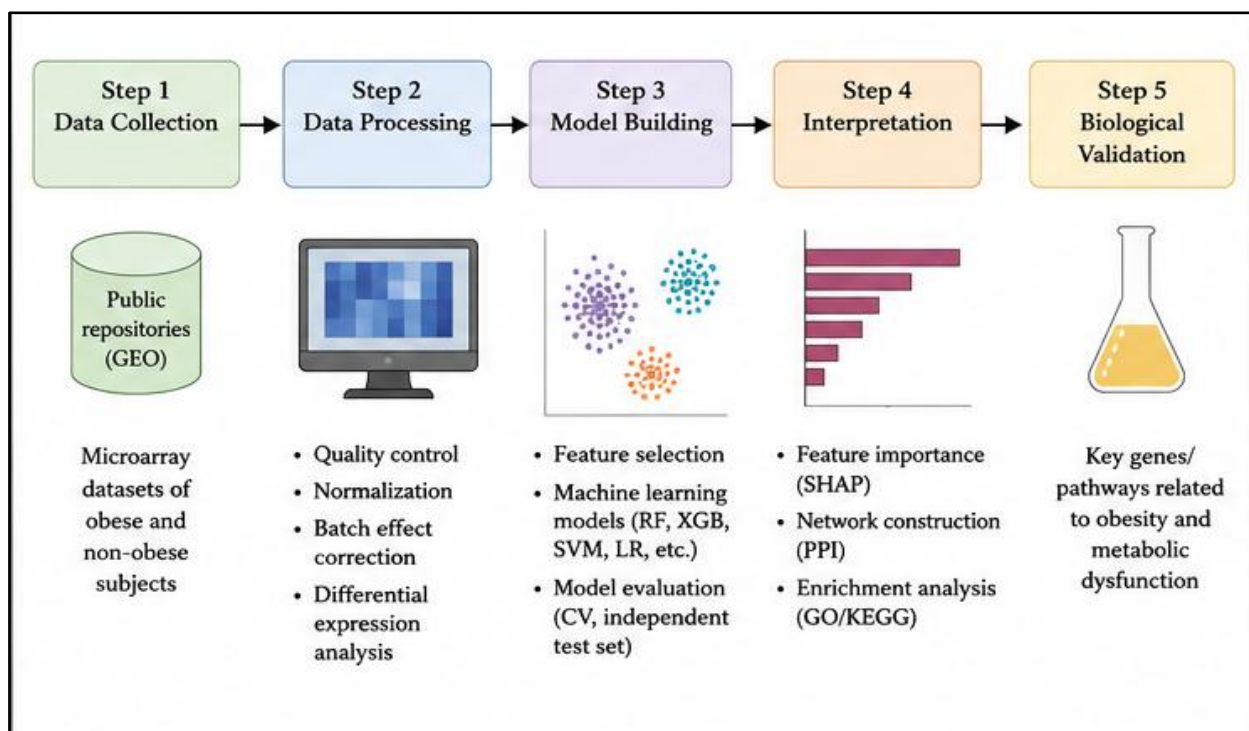


Fig.1 Workflow of the study

2.2 Data Preprocessing and Transcriptomic Integration

Transcriptomic datasets were subjected to preprocessing to enhance the biological consistency and minimize technical variation and to ensure the consistency of data from datasets collected under different experimental conditions and platforms [33,38]. Systematic pre-processing of the transcriptomic data is crucial to obtain reliable and biologically meaningful computational models, because many datasets are often contaminated with platform-specific effects, missing expression values and inter-sample variations. In the present study the preprocessing pipeline used was the gene filtering, missing value imputation, normalization and batch effect correction followed by machine learning implementation, all of which were done previously to the implementation of machine learning [35,40].

Transcriptomic expression matrices for GSE22309 and GSE103766 were downloaded from the GEO repository, as were the platform annotation files [35]. GEO platform annotation files include detailed information on probe identifiers, gene symbols and transcript annotations necessary for the biological interpretation of expression profiles [35].

The two GEMS GSE22309 and GSE103766 were collected from different tissue types and different experimental platforms, so direct integration without feature harmonization would create inconsistencies and bias based on these platforms [38]. Thus, the common genes that were represented in both sets were used to generate a common gene transcriptomic expression matrix [37]. Common genes were retained, leading to biological comparisons and the identification of conserved obesity-associated molecular signatures in adipose tissue and in skeletal muscle [28]. However, transcriptomic data often are missing expression values because of the variability of expression at the level of the probe, inconsistencies in the signal, or noise in the experiment [39].

The idea of normalizing is especially relevant for transcriptomic data sets where there are thousands of molecular variables that are all heterogeneous [39].

The selected datasets were derived from various biological experiments and could be from various microarray platforms, so a batch effect correction was also performed to reduce non-biological technical variation [38]. Systematic differences that are observed in the experimental conditions, laboratory procedures, or platform-specific biases are considered batch effects and are not biological differences [99].

After preprocessing, normalization, and batch effect correction, an integrated transcriptomic matrix of the biologically comparable gene expression features was generated to be used for downstream machine learning implementation and EAI analysis [40]. The resulting expression matrix was used as the basis for discovering obesity-associated transcriptomic biomarkers using an ensemble learning and systems biology approach [44].

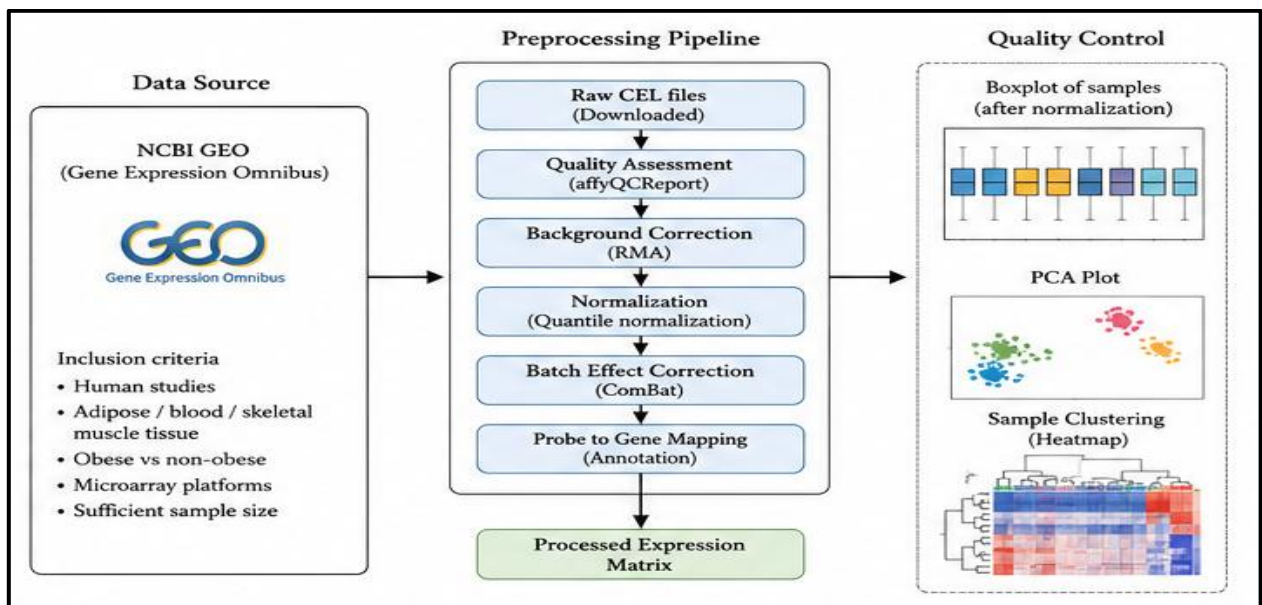


Fig.2 Acquisition and Preprocessing of Gene Expression Data

2.3 Construction of Feature Matrix and Target Variable

The integrated gene expression data were then preprocessed and normalized for transcriptomics and converted into a machine learning-friendly format [40]. In classification problems, machine learning models need to be trained with a set of input features (predictor variables) and output labels (target variables) in a structured format of numbers [78]. For this reason, a feature matrix and target variable were created to enable supervised machine learning analysis of the obesity-associated transcriptomic profile [79].

In order to obtain a common feature matrix (X) between GSE22309 and GSE103766, the integrated transcriptomic expression matrix with normalized common genes was used [40]. Here the rows represented biological samples and the columns gene expression features for fixed transcriptomic variables [33].

The target variable (Y) was outlined by obesity associated phenotype labels [40]. Samples of healthy, or metabolically normal individuals were classified in the control class (0), and samples of those with obesity-related phenotypes, such as insulin resistance or metabolically abnormal obesity states, were classified in the obesity class (1) [38]. A binary phenotyped assignment was used to support supervised classification modelling to differentiate obesity-associated transcriptomic signatures from control expression profiles [79].

The feature matrix and target variable were built and supervised machine learning algorithms were implemented to identify the nonlinear patterns of transcriptomic changes associated with obesity [41]. This type of structuring also enabled the downstream training, validation and SHAP-based explainability analysis of these models.

2.4 Machine Learning Model Development

After the preprocessing of the transcriptomic data and construction of the feature matrix, supervised machine learning models were applied to uncover obesity-related transcriptomic signatures and to classify biological samples from gene expression profiles [40,41]. Machine learning methods were chosen due to their effectiveness in handling high dimensional biological data and detecting complex nonlinear molecular interactions that are often missed by traditional statistical techniques [39]. There are thousands of genes in the obesity-associated transcriptomic data, and the relationships between these genes are highly complex, so that machine learning algorithms can be used to provide a computational framework for biomarker discovery and disease

Dataset splitting was used to train the model with 80% and to test it independently with 20% of the data samples [40]. In order to avoid sampling bias and ensure balanced class representation, stratified sampling was used, resulting in a training and testing set of the same size and balanced distribution of obesity and control phenotypes [81]. Furthermore, a fixed random state parameter was used to guarantee the computational results' reproducibility and consistency [78].

First, several supervised machine learning classifiers were tested for their classification performance and to find the best one for obesity transcriptomic analysis [41].

Logistic Regression is a method that estimates probabilities of classes using a sigmoid function and is widely used for biomedical classification problems that have binary classes [84]. In omics-scale analyses, however, transcriptomic data are often highly-dimensional and can have nonlinear relationships between genes, which can lead to poor predictive performance of Logistic Regression [39].

A Decision Tree (DT) classifier was also developed due to its easily explainable rule-based classification function [86]. The Decision Trees recursively partition features based on decision

thresholds and offer explainable classification routes [87]. When the biological data are noisy or high dimensional, however, individual decision trees are prone to overfitting and instability [88].

Support Vector Machine (SVM) models were used because they optimise hyperplane parameters and transform the data into kernels, which are able to classify complex biological data sets [84]. SVM has proven to be very useful in transcriptomic classification problems, which involve the presence of high dimensional molecular variables [85]. When used with highly heterogeneous transcriptomic data with large feature spaces, however, computational efficiency and interpretability can become a challenge [39].

K-Nearest Neighbors (KNN) algorithm as a distance-based classification method, where nearest transcriptomic profiles are used to classify the sample [82]. KNN has been used for disease classification; in these studies, dimensionality and sensitivity to irrelevant transcriptomic variables can affect model performance [83].

In addition, Gaussian Naïve Bayes classification was tested due to its use of a probabilistic learning method using Bayes' theorem [85]. Although it is simple and computationally efficient, Naïve Bayes makes the assumption of feature independence, which might limit the classification accuracy in transcriptomic datasets where biological variables are not independent [39].

Random Forest was chosen as the main classification algorithm because it is robust and can capture the non-linear relationships among the transcriptomes [90]. Random Forest is based on the principle of bootstrap aggregation (bagging) and random feature selection, which are used to generate a set of decision trees and minimize variance and enhance the stability of classification [91]. Random forest has proven to be useful in biomarker discovery studies because of its ability to rank the importance of transcriptomic features and its ability to efficiently process a large number of genes [92].

To enhance predictive robustness the number of decision-trees in the Random Forest classifier was optimized to 500 (n estimators = 500) [91]. The more trees in the model, the more stable the model and the less variability in the classification, and the more consistent the prediction [90].

In parallel, XGBoost was used as an ensemble classifier based on gradient boosting to make a comparison in terms of predictive power [93]. XGBoost is iterative, and corrects classification errors made by previous learners in a sequential manner [94]. In omics-based disease classification problems, XGBoost often outperforms traditional algorithms in terms of accuracy and reliability [95].

Comprehensive performance comparison was enabled by implementing several supervised learning algorithms, and the most biologically valid and computation efficient framework for downstream explainability analysis and identification of obesity biomarker was selected for implementation [44].

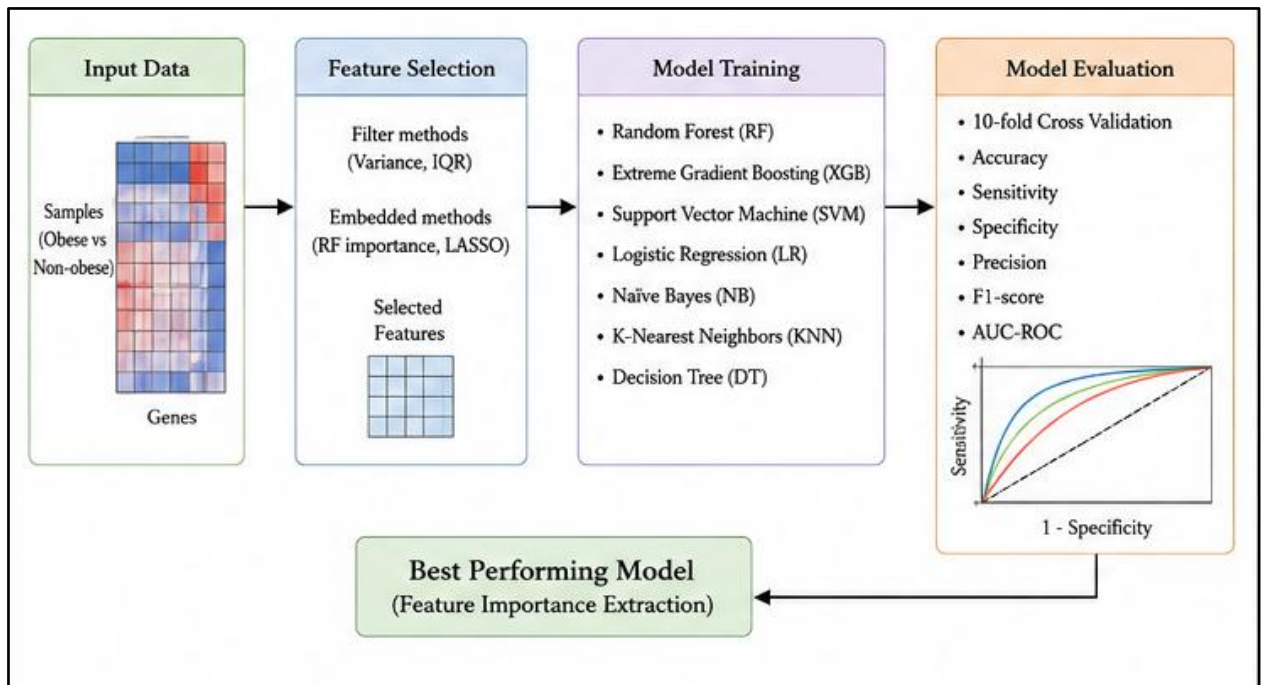


Fig.3 Machine Learning Pipeline for Classification

2.5 Model Evaluation and Performance Assessment

Model evaluation is a crucial part of supervised learning as sometimes the predictive accuracy of the model does not necessarily align with the biological reliability and classification robustness of the model, especially in biomedical datasets with class imbalance and transcriptomic heterogeneity [81]. Thus, several statistical performance measures were used to thoroughly evaluate classification ability and generalizability of the model [41].

Model performance was evaluated on the independent test dataset (20%) which was reserved during train–test splitting [40]. The use of an unseen test set allowed to objectively evaluate predictive generalization, and to reduce the risk of over-fitting, where models learned the training set and not the biologically meaningful transcriptomic patterns [88]. Transcriptomic datasets are characterized by high dimensional features and also by few biological samples and this is why independent validation is of paramount importance in order to guarantee computational reliability [39].

These metrics give complementary information about classification behavior and taken together the information about the predictive performance in the context of obesity-associated transcriptomic analysis is enhanced [41].

Precision measures the percentage of correct predictions of obesity related samples out of all positive predictions [79]. The higher the precision value, the more reliable the classification of obesity predictions are and the lower the number of false-positive predictions [81].

Good recall values mean that there are fewer false negative classifications and that the obesity-associated transcriptomic profile is effectively identified [81].

AUC values close to 1 are indicative of good predictive discrimination and values close to 0.5 are similar to random classification [79].

Comparative evaluation of machine learning models helped identify the best algorithm for downstream explainability analysis and prioritizing transcriptomic biomarkers [44]. Ensemble learning methods often outperform traditional classifiers for high dimensional omics data sets, so special focus was given to these models: Random Forest and XGBoost [41].

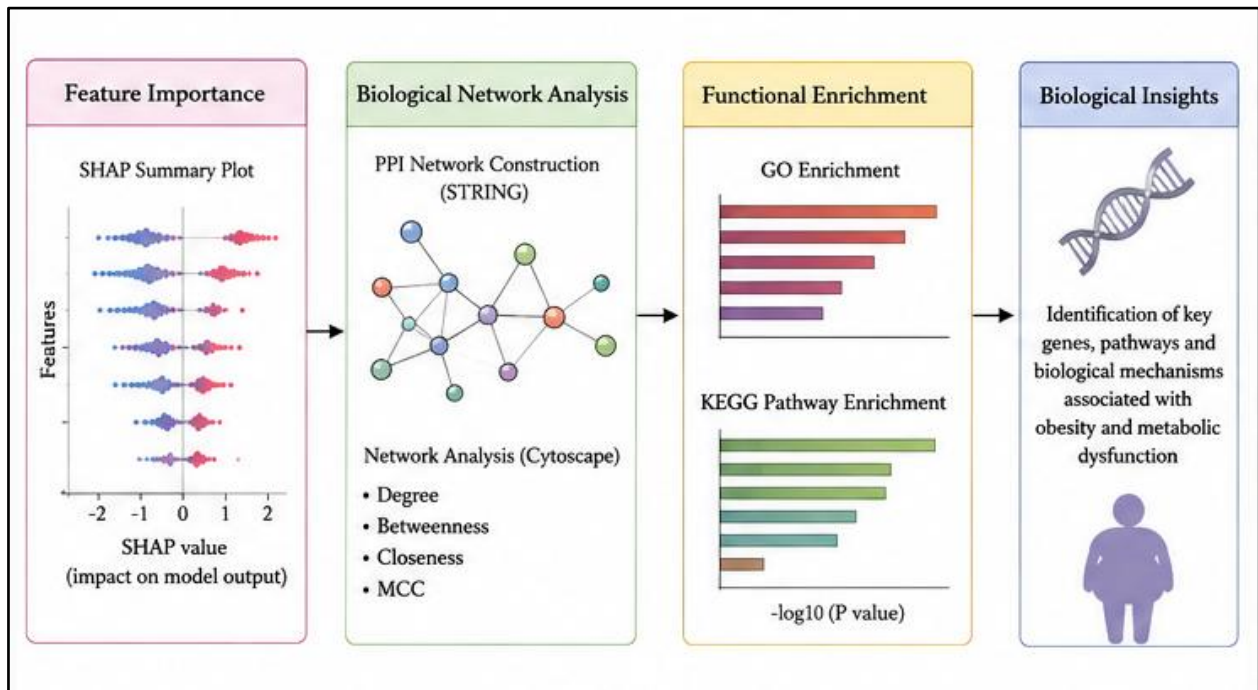


Fig.4 Post Modelling Analysis and Biological Interpretation

2.6 Explainable Artificial Intelligence (XAI) Analysis with SHAP

Machine learning model assessment was followed by SHAP analysis to gain insight into the interpretability of the predictive outcomes and to define the transcriptomic features that play significant role in obesity classification [44]. While ensemble machine learning models can achieve high performance in prediction, their decision-making process is still challenging to interpret due to their black-box nature [42]. For this reason, explainable AI techniques are growing in significance for biomedical research, enabling biological interpretation and enhancing the trust in computational forecasts [43].

SHAP offers an extra advantage over the conventional feature importance methods in that it offers global interpretability (average importance of the features for all the predictions) and local interpretability (explanation of the behavior of the model at a specific prediction level) [101]. This interpretability is especially useful in the context of transcriptomic obesity studies, where the ability to prioritize potential obesity biomarkers in a biologically sensible way becomes easier [45].

TreeExplainer efficiently computes exact Shapley values, and is also computationally efficient in high-dimensional biological datasets [103]. The thousands of gene expression variables in transcriptomic data made computational scalability an advantage of TreeExplainer that made it ideal for obesity-associated biomarker discovery [44].

Genes with high SHAP values were deemed to be biologically significant transcriptomic features that played a significant role in discriminating between the obesity and control phenotypes [45]. Global SHAP analysis allowed to prioritize obesity-associated molecular biomarkers for downstream biological interpretation [44].

To gain further insight into the contributions of transcriptomic features across biological samples, summary plots and beeswarm plots of SHAP values were created [103]. SHAP summary plots show the distribution of feature importance by arranging transcriptomic variables in order of their 'global' predictive contribution [101]. Beeswarm plots also show the direction of gene expression change between samples (up or down), and whether this change has a positive or negative association with obesity classification [102].

For the local explanations, force plots and waterfall plots were used to produce local-level SHAP explanations [103]. These visualisations are used to elucidate prediction results for specific biological samples based on the cumulative contribution of transcriptomic features that inform

the prediction [44] These local interpretability methods help to promote understanding of the specific effects of genes on obesity-related transcriptomic predictions [101].

Biologically interpretable prioritization of transcriptomic biomarkers associated with obesity was achieved by integration with machine learning classification of SHAP [45]. The prioritized genes obtained by SHAP were then further analyzed using systems biology methods such as protein–protein interactions and pathway enrichment analyses for their biological significance [46].

2.7 Protein–Protein Interaction Network Analysis.

The protein–protein interaction (PPI) network analysis was then conducted to explore functional interactions between the transcriptomic biomarkers identified by machine learning and SHAP analysis [46,104]. Genes do not always act alone but work together through molecular pathways, so PPI network analysis can be used to find functional relationships among candidate biomarkers [105]. These systems biology approaches help to deepen the understanding of the underlying mechanism of disease-associated transcriptomic signatures and aid in prioritizing important genes in biologic pathways that contribute to pathogenesis of obesity [106].

Downstream PPI network construction was performed after induction of the transcriptomic biomarkers using feature importance analysis based on SHAP [44,45]. SHAP selects genes that make a large contribution to the obesity classification, which allows for better confidence in their biological relevance and to interpret their functions through interaction network analysis [46].

STRING is broadly used for transcriptomic and systems biology research since it generates extensive protein interaction networks from known or predicted protein interactions [111]. The

candidate genes were added to the STRING database to obtain networks and find proteins that are biologically connected to obesity [112].

Confidence thresholds for interaction were used to prune false-positives interactions and to guarantee biological reliability of interaction networks generation [111]. For downstream analysis, only statistically and biologically relevant interactions supported by experimental evidence, co-expression data and databases were considered [112]. This facilitated the confidence in identified obesity associated molecular interactions and minimized the inclusion of biologically irrelevant interactions [46].

The STRING interaction network was then exported and analyzed with Cytoscape for the in-depth visualization and topological analysis of the network [113]. Cytoscape is a popular systems biology tool to visualize molecular interaction networks and to detect central regulatory nodes [114]. PPI networks were visualized to examine connectivity patterns among obesity associated biomarkers, and to better interpret the molecular relationships [115].

The biologically important genes in interaction networks were identified by applying topological network analysis techniques based on centrality measures [114]. Genes with high interaction connectivity were deemed as hub genes as they often are involved in several biological pathways and significantly in the molecular mechanism associated with diseases [109].

Hub genes led to the prioritization of transcriptomic biomarkers likely to be linked to obesity-associated metabolic dysregulations, mitochondrial impairments and inflammatory signaling [106]. Hub genes also offer a mechanistic understanding of obesity pathophysiology, and may be targeted for therapeutic intervention or precision medicine applications [109]. This led to better biological validation of machine learning prioritized transcriptomic signatures [46].

2.8 Functional Pathway Enrichment Analysis

Functional pathway enrichment analysis was conducted to assess the biological relevance of obesity-associated transcriptomic biomarkers that were discovered using machine learning and SHAP analysis [47]. Often, identification of a transcriptomic biomarker results in several potential target genes, which can have complicated relationships, which makes it important to use enrichment analysis for systematic interpretation of their role in molecular pathways and cellular processes [46]. Functional enrichment thus offers mechanistic insights of obesity-associated biological changes and confirms transcriptomic results in known metabolic contexts [116].

Candidate genes identified by SHAP and PPI network analysis were analyzed for pathway enrichment by the online functional annotation platform Enrichr [119] that integrates several biological pathway databases. Enrichr enables the identification of statistically enriched biological pathways, by comparing sets of candidate genes with curated molecular databases [120]. This strategy allowed the systematic study of molecular processes over-represented in obesity-associated transcriptomic biomarkers [121].

In addition to molecular pathway maps, KEGG offers curated maps related to metabolism, cellular signalling, disease progression and physiological regulation [116]. Reactome also provides detailed pathway annotations for molecular interactions and biological reactions involved in cellular homeostasis as well as disease processes [118]. Integration of both databases resulted in biochemical and biological validation of the enrichment analysis [47].

Adjusted p values and enrichment scores were calculated using Enrichr [120] to determine statistical enrichment significance. Candidate genes that showed statistically significant enrichment in the pathways were selected for further biological interpretation [119]. This

strategy allowed the identification of molecular pathways that could be involved in metabolic dysfunction and transcriptomic dysregulation associated with obesity [121].

Pathways expected to show significant enrichment based on previous obesity transcriptomic data included those involved in oxidative phosphorylation, reactive oxygen species signaling, insulin resistance, fatty acid metabolism, glucose homeostasis, and inflammatory signaling [36,117]. The identification of enrichment in these pathways would further support biological relevance of machine learning-prioritized biomarker [24,29].

Functional enrichment analysis also enabled the biological interpretation of transcriptomic signatures by connecting computational results with known mechanisms that are associated with obesity [47]. Genes that are enriched in the mitochondrial and metabolic pathways may be related to impaired oxidative metabolism, while genes in inflammatory pathways may be related to regulatory functions in chronic metabolic inflammation [121]. This translation of biological relevance and significance of transcriptomic biomarkers at the pathway level greatly enhances biological understanding and translational meaning of these biomarkers at the pathway level [46].

The integration of machine learning, SHAP explainability, protein interaction analysis and pathway enrichment led to the development of a biologically interpretable and systems-level computational framework for obesity-associated biomarker discovery, respectively [44,46]. This integrative approach not only deepened the understanding of the mechanistic aspects of obesity progression but also boosted the reliability of transcriptomic biomarker prioritization [98].

CHAPTER 3

RESULTS

3.1 Implementation of Machine Learning Models and Comparative Performance Evaluation

Transcripts were preprocessed and models were built, followed by supervised machine learning algorithms for classification of obesity-associated transcriptomic profiles using integrated gene expression data from GSE22309 and GSE103766 [35,40]. Various classification models were tested to find the most reliable computational framework to discriminate an obesity-associated molecular signature from control transcriptomic profiles [41]. Comparative model assessment was necessary to find biologically sound predictive methods, due to the high dimensionality and nonlinear molecular interactions of obesity-associated transcriptomic datasets [39].

Systematic analysis of predictive performance using these classifiers compared their performance using traditional statistical classifiers, tree-based classifiers and ensemble learning classifiers [41].

The results of performance assessment revealed significant differences in performance across different classifiers implemented, which were attributed to varying capability of the classifiers to model the nonlinear transcriptomic relationships and high dimensional biological features [39]. The predictive performance of traditional statistical and probabilistic models like Logistic Regression and Gaussian Naïve Bayes was comparatively lower due to their limited capacity to capture the complex transcriptomic interactions [83,85]. Likewise, Decision Tree and KNN classifiers had moderate predictive power, but they were not as reliable or robust in highly heterogeneous biological datasets as compared to other classifiers [82,88].

Ensemble learning methods were the best performing algorithms across the group, showing that they are more suitable for obesity-associated transcriptomic analysis than the other evaluated algorithms [41]. The classifiers outperformed traditional classifiers on most statistical performance measures, as they were able to learn nonlinear relationships, with less overfitting and greater predictive stability [89,93].

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	0.79	0.80	0.79	0.79
Logistic Regression	0.76	0.76	0.76	0.76
SVM	0.56	0.62	0.56	0.50
Naïve Bayes	0.67	0.70	0.67	0.67
Random Forest	0.74	0.75	0.75	0.75
KNN	0.62	0.63	0.63	0.62

TABLE I. Classification Performance of Machine Learning Models

The results observed here indicate that machine learning algorithms with the ability to model nonlinear feature interactions [39] are very useful for the classification of obesity, based on transcriptomic data. Ensemble learning methods, specifically Random Forest and XGBoost, outperformed other algorithms in finding obesity-associated transcriptomic signatures due to their noise resistance and their resistance to dimensionality issues [89].

3.2 Analysis of the Confusion Matrix of Random Forest Classifier

A machine learning model of random forest with 500 decision trees ($n_estimators = 500$) was chosen for detailed evaluation [90,91]. To evaluate the prediction outcomes and quantify the classification errors, further evaluation was carried out using the Random Forest, the balanced

predictive performances with better classification stability was achieved, therefore the further evaluation was carried out using confusion matrix analysis [79].

The optimized Random Forest classifier showed overall classification accuracy of about 74.63% on the independent testing dataset, suggesting the classifier had good prediction accuracy for discriminating obesity-associated transcriptomic profile from control cases [40]. The classification accuracy observed indicates the ability of ensemble learning techniques to deal with the high level of transcriptomic complexity and with the nonlinear molecular interactions that are often encountered in obesity data [41].

A confusion matrix was created to represent the prediction results and assess the discrimination of the model between obesity and control transcriptomic phenotypes [79]. The confusion matrix analysis indicated that the Random Forest (RF) classifier had appropriate classification rates with a significant number of obesity and control samples as well [81].

Classification Outcome	Number of Samples
True Positive (TP)	25
True Negative (TN)	27
False Positive (FP)	7
False Negative (FN)	8

Table 2 Confusion Matrix Interpretation of Random Forest Model

These results suggest that misclassification rate is relatively low and the prediction capability is fairly balanced among transcriptomic classes [79].

The precision, recall and F1-score values also helped to validate the Random Forest (RF) classifier [79]. This well-balanced predictive performance is especially relevant in biomedical

classification applications where mispredictions could impact biological interpretation and biomarker prioritization [44].

Differential molecular patterns associated with obesity also proved to be reproducible when applied to the confusion matrix with machine learning models [39]. Random Forest correctly classified the transcriptomic profiles, which indicates the presence of biologically informative molecular signatures, as obesity is a coordinated dysregulation of mitochondrial metabolism, inflammatory signaling and glucose homeostasis pathways.

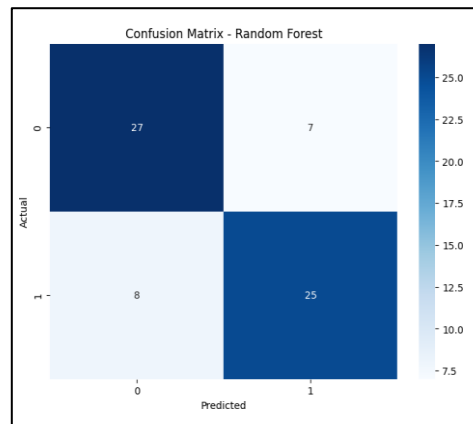


Fig. 5 Random Forest confusion matrix showing robust classification performance between control and obesity samples

The observed predictive performance led us to choose the Random Forest model for downstream SHAP-based explainability analysis, from which we identified a number of transcriptomic biomarkers that were significant for the classification of obesity and biologically explained [44].

3.3 SHAP Based Global Feature Importance Analysis

SHAP was particularly beneficial since the ensemble models often operate as black-box classifiers. This explainability is especially valuable for transcriptomic studies as it can help with biologically relevant biomarker prioritization and increase the trust in computational predictions [45].

The initial aim of global SHAP analysis was to estimate the contribution of transcriptomic features on average over all the biological samples [102]. Genes with higher values of SHAP contribution were interpreted as having a greater effect on transcribing obesity-associated phenotype from a control phenotype [44].

A small number of transcriptomic biomarkers showed disproportionately high contributions to predictions of the model in the SHAP global importance analysis [102]. A number of high-ranking genes were linked to mainly mitochondrial metabolism, oxidative phosphorylation, RNA regulation, and cytoskeletal organization, indicating a significant role for these biological pathways in obesity-associated metabolic dysregulation [24,36].

Genes with strong predictive contribution to obesity classification among the top ranked transcriptomic biomarkers were the BDH1, MRPS31, RARS2, STOML2, AKAP10, UCP3, NUCB2, SRSF1 and ACTG2 genes [40]. These genes are involved in metabolic and mitochondrial functions that are important for obesity pathogenesis [24].

Likewise, RARS2 and SRSF1 are related to RNA processing and mitochondrial translational control, indicating possible changes in transcript levels that could impact cellular energy metabolism and protein production [123]. In addition, ACTG2, a cytoskeletal organisation-related gene, contributed significantly to classification results, signifying potential changes in structural remodeling and metabolic tissue organisation associated with obesity [125].

One of the key features of the global SHAP analysis was the sharp drop in feature importance after the top-ranked biomarkers, suggesting that there is a relatively clustered obesity-associated transcriptomic signature [102]. This observation implies that although transcriptomic-obesity association datasets are highly dimensional, a limited subset of genes may have disproportionately large control over the classification results [38]. This focus of predictive importance greatly enhances the biological interpretability, and also helps prioritized biomarker identification [44].

The resulting SHAP feature importance bar plot showed ranked transcriptomic contributions which allowed visualization of the genes with the strongest effect on obesity classification outcomes [101]. This analysis also confirmed that machine learning models could detect biologically relevant transcriptomic signatures linked to metabolic dysfunction that is associated with obesity [45].

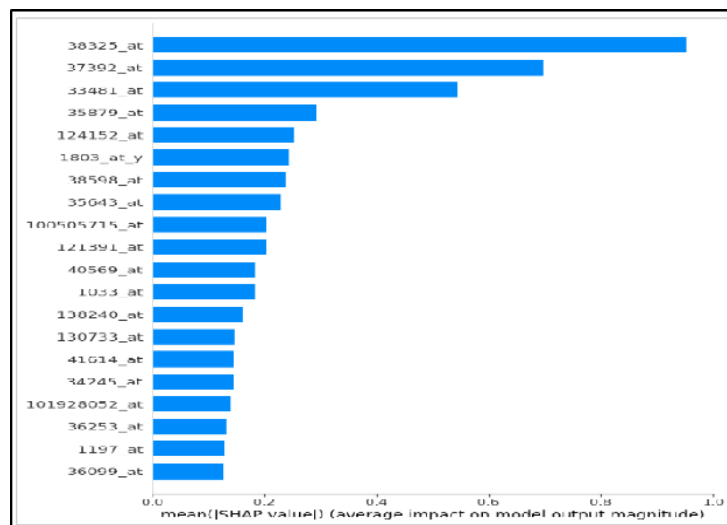


Fig. 6 SHAP global feature importance bar plot highlighting the top 20 obesity-associated transcriptomic biomarkers.

3.4 Cohort-Level SHAP Interpretation of Transcriptomic Features

At the cohort level, feature importance can be calculated to evaluate the effect of the variation of a single gene's transcriptomic expression on the classification results of the obesity phenotype [102]. This kind of analysis can help shed light on the biological heterogeneity and molecular variability linked to obesity [38].

A beeswarm plot was created by the transcriptomic feature contribution to each sample in the study [103]. Each point in the beeswarm plots includes a single biological sample and the associated value of the SHAP contribution for a given gene [101].

The SHAP beeswarm analysis revealed distinct obesity and control phenotype expression states [44]. There was a substantial directionality effect for a number of top-ranked genes, indicating their key role in the molecular dysregulation involved in obesity [45]. High-ranking transcriptomic biomarkers showed wide distributions of SHAP values, indicating that these values contributed to the prediction heterogeneity of biological samples across a broad range [101].

Genes involved in metabolism of the mitochondria and oxidative phosphorylation proved to have different expression-dependent effects on obesity classification [73], such as genes BDH1, STOML2, MRPS31, and UCP3.

Further, some transcriptomic biomarkers related to RNA processing and metabolic signaling, such as RARS2, SRSF1, and NUCB2, showed strong directionality of their influence on classification predictions [123,124]. These observations highlight the complex reconfiguration of multiple molecular pathways that are associated with obesity, rather than a single disturbed pathway [36].

Obesity is highly phenotypically and molecularly heterogeneous, so the contribution of different transcriptomic features was expected to vary and further underscores the significance of interpretable machine learning models to discover obesity biomarkers [38].

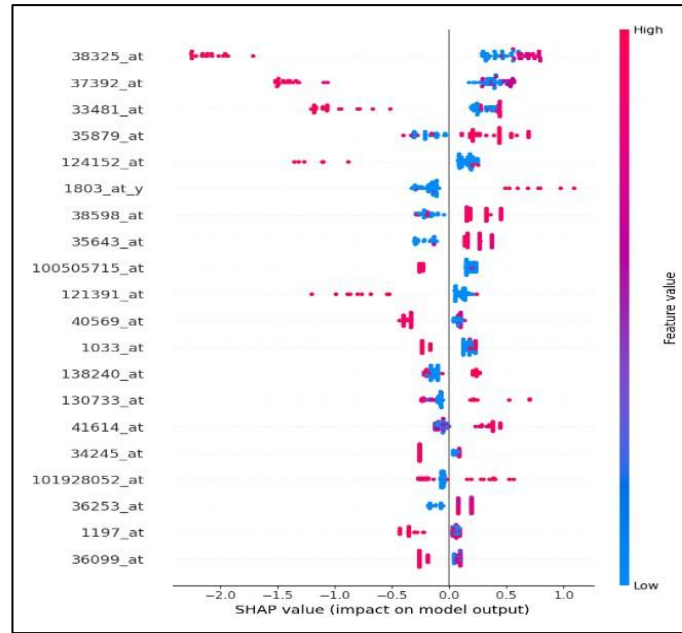


Fig. 7 SHAP beeswarm plot illustrating cohort-level distribution and directional contribution of top-ranked transcriptomic features.

Overall, SHAP interpretation at the cohort level confirmed the relevance of the identified transcriptomic biomarkers and offered a mechanistic view of the role that the alterations in the expression of these genes, as a group, play in the classification results [44]. These results also helped downstream local interpretation analysis and systems biology validation of prioritized obesity-associated transcriptomic signatures [46].

3.5 Local Interpretability of Obesity Classification Using SHAP

Local SHAP analysis was carried out to explore the transcriptomic contributions affecting obesity classification at the individual sample level [101]. Global and cohort-level SHAP analyses give an overview of the importance of a transcriptomic data set and population-level trends, respectively, while local interpretability allows to study how combinations of genes affect the prediction results of specific biological samples [44]. This is especially relevant for obesity research, as obesity is a highly molecular- and metabolically heterogeneous disease [38]. Waterfall plots and force plots of SHAP values [103] were used to perform local interpretability analysis using the plots of cumulative contribution of transcriptomic features to the final model predictions. These plots help to quantify the positive and negative effects of individual genes on classification results and help to visualize the effect of the transcriptomic features on predictions as they shift toward the obesity or control phenotypes [101]. Genes that contribute positively to the classification of obesity tend to favor the prediction of obesity transcriptomic state, while genes that contribute negatively to obesity tend to favor the prediction of non-obesity transcriptomic state [102].

Several transcriptomic markers worked together to impact prediction results and demonstrated a multifactorial aspect of molecular obesity-related metabolic dysfunction [10,36]. These results further emphasize the value of ensemble learning frameworks in discovering non-linear transcriptomic relationships that are not captured by traditional statistical techniques [39].

Genes such as NUCB2 and SRSF1 showed significant and positive association with obesity classification in representative samples among the most influential local transcriptomic contributors [124]. These genes were found to have high SHAP values indicating their possible role in metabolic deregulation and disruptions in endocrine signaling during obesity [17].

NUCB2 has been previously associated with the regulation of appetite and metabolism, which further underscores the importance of its role in obesity-related mechanisms [124].

By contrast, a number of mitochondrial-associated genes, such as MRPS31, STOML2, BDH1, and UCP3, showed negative values in some biological samples, suggesting transcriptomic changes that may be linked to the disruption of mitochondrial metabolism and decrease in oxygen capacity [24,73].

The overall contribution of the transcriptomic biomarkers produced unique prediction scores for each biological sample and demonstrated transcriptomic heterogeneity of obesity-associated molecular phenotypes [38]. Cumulative SHAP contributions for representative obesity associated predictions showed prediction shifts towards obesity class, whereas control samples showed opposing transcriptomic contributions [101]. These data suggest that obesity classification is enabled by an integrated molecular dysfunction rather than by single gene expression variations [36].

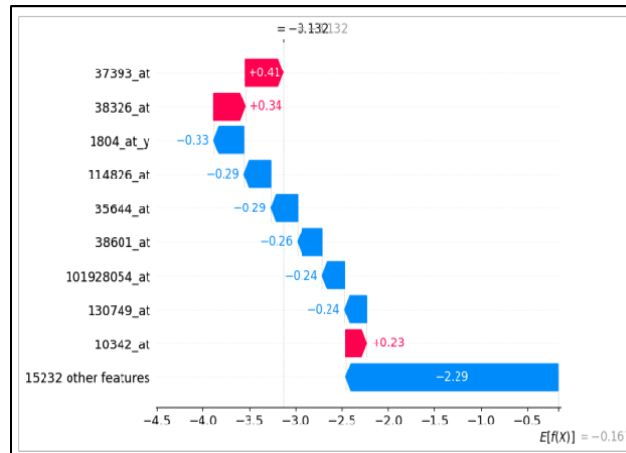


Fig. 8 SHAP waterfall plot showing local feature contributions driving an individual obesity prediction.

The force plots from SHAP further depicted the directionality of the influence of the transcriptomic biomarkers on the prediction results by showing the cumulative effect on pushing

towards obesity and control classifications [103]. These local explainability strategies greatly enhanced the interpretability of machine learning outcomes and allowed for a biologically meaningful understanding of obesity-associated transcriptomic heterogeneity [44].

In aggregate, the local SHAP interpretability results showed that coordinated transcriptomic interactions that involve mitochondrial metabolism, regulation of RNA, remodeling of the cytoskeleton and/or endocrine signaling had an effect on obesity-associated classification outcomes [24,36]. These results further validated the machine learning predictions, and also provided biologically interpretable evidence to support identified obesity-related biomarkers [45].

3.6 Protein–Protein Interaction Network Analysis of Identified Biomarkers

Many transcriptomic biomarkers are part of networks of genes that regulate various aspects of disease pathogenesis, and network analysis helps identify relevant genes for biological processes [104].

The top biomarkers identified using SHAP were loaded into the STRING database to obtain experimentally confirmed and predicted molecular interactions [110]. The interaction network they found revealed significant functional connectivity between the obesity-associated genes identified, suggesting co-ordinated molecular regulation of metabolic and inflammatory processes [111]. By topological network analysis, NUCB2, SRSF1 and ACTG2 were identified as well connected hub genes with high network connectivity, performed using Cytoscape software [113,114]. Hub genes are highly interconnected molecular nodes that are often targeted in the regulation of several downstream biological pathways and disease-associated processes [107]. Therefore, the identification of these genes suggests their possible role in obesity-related metabolic dysfunction and gene expression regulation.

NUCB2 was shown to be a hub gene with high connectivity and biological relevance because of its prior reported role in appetite regulation, endocrine signaling and metabolic homeostasis [124]. The dysregulation of NUCB2 might thus be implicated in the energy imbalance and endocrine disturbances associated with obesity [17]. Changing pathways involved in RNA processing could affect the stability of mitochondrial transcripts and metabolism involved in obesity [36].

ACTG2 is mainly involved in cytoskeletal organization and structural remodeling was also identified as a biologically important hub gene [125]. ACTG2 interactions indicated potential tissue remodeling and metabolic changes with obesity that could be addressed in skeletal muscle tissues [22].

Transcriptomic data also showed strong associations with the genes of mitochondrial metabolism and oxidative phosphorylation, which were further supported by network visualization [24,73], supporting a role of mitochondrial dysfunction in obesity progression [26].

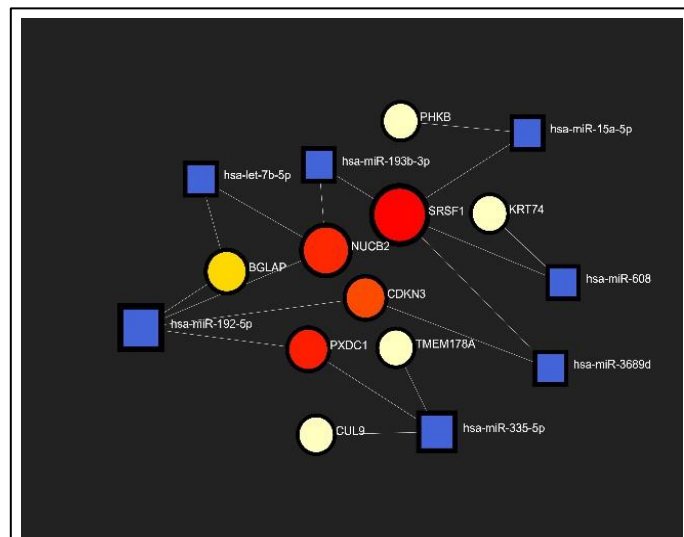


Fig. 9 STRING–Cytoscape protein–protein interaction network identifying NUCB2, SRSF1, and ACTG2 as central hub genes.

PPI network analysis greatly enhanced biological confidence in machine learning-prioritized biomarkers by showing their involvement in coordinated molecular pathways involved in obesity pathophysiology [46].

3.7 Functional Pathway Enrichment Analysis of Obesity-Associated Transcriptomic Biomarkers

This systems biology approaches also boost the reliability of computational predictions by testing candidate genes in biologically meaningful pathways [46].

Priority transcriptomic biomarkers from the SHAP analysis and the protein–protein interaction network were then put through an enrichment analysis to determine the biological pathways that are statistically overrepresented [119]. Functional enrichment identified metabolic pathways that are heavily implicated in obesity-associated metabolic dysfunction including pathways for oxidative phosphorylation, reactive oxygen species (ROS) signaling, glycogen metabolism, insulin resistance and mitochondrial bioenergetics [24,36].

One of the most remarkable biological processes that appeared among the significantly enriched pathways was oxidative phosphorylation, which was among the identified transcriptomic biomarkers [117].

Other genes involved in insulin resistance pathways also showed significant enrichment [117]. Thus, the enrichment of insulin resistance pathways further confirmed their biological relevance of the machine learning prioritized transcriptomic biomarkers [31].

Other pathways related to glycogen metabolism and glucose homeostasis were also also significantly enriched [121]. Impaired mechanisms in energy regulation in obesity associated transcriptomic profiles are hence supported by dysregulated glycolytic metabolic pathways [22].

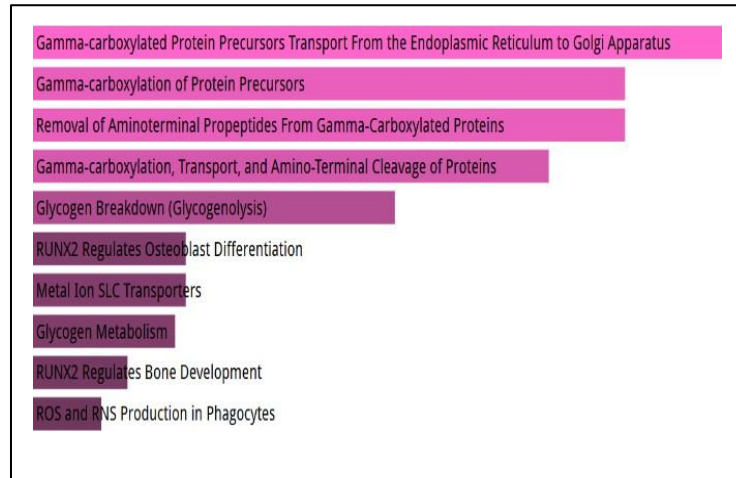


Fig. 10 Enrichr KEGG/Reactome pathway enrichment analysis revealing mitochondrial dysfunction and metabolic dysregulation pathways in obesity.

Reactome-based enrichment also showed pathways related to mitochondrial organization, ATP synthesis, electron transport chain activity and cell stress response pathways [118]. The results are well correlated with transcriptomic data that shows mitochondrial metabolic impairment is a central mechanism of obesity [24]. The biological relevance of genes with high SHAP contribution values during transcriptomic classification [73] such as BDH1, MRPS31, STOML2 and UCP3 is further supported by the mitochondrial pathways they are enriched in.

Pathway enrichment analysis revealed strong relationships between obesity-associated transcriptomic biomarkers identified by machine learning and pathways of mitochondrial dysfunction, oxidative stress, insulin resistance and metabolic dysregulation [24,36].

Pathway	Biological Relevance in Obesity
Oxidative Phosphorylation	Reduced ATP production and mitochondrial dysfunction
ROS/RNS Signaling	Oxidative stress and inflammation
Glycogen Metabolism	Altered glucose utilization and insulin resistance
Insulin Resistance	Impaired glucose uptake and metabolic dysfunction
Mitochondrial Metabolism	Reduced fatty acid oxidation and metabolic flexibility

Table 3 Major Enriched Biological Pathways Identified in Obesity

3.8 Integrated interpretation of transcriptomic findings

The machine learning framework was able to achieve a balance predictive performance to distinguish obesity-associated molecular profiles from control samples, and successfully identify transcriptomic features capable of doing so [41].

The same explainability analysis methods used for SHAP were employed to further increase the interpretability of the transcriptomic predictions by highlighting the genes that have a significant contribution to the classification of obesity that are biologically relevant [44]. A number of biomarkers were identified, such as BDH1, MRPS31, RARS2, STOML2, AKAP10, UCP3, NUCB2, SRSF1 and ACTG2, which revealed significant role of mitochondrial metabolism, RNA regulation and cytoskeletal remodeling in obesity pathophysiology [40]. The results confirm emerging evidence that points towards mitochondrial dysfunction and metabolic inflexibility as central players in the progression of obesity [24].

Protein–protein interaction analysis further confirmed the biological confidence of prioritized biomarkers by identifying highly connected hub genes, such as NUCB2, SRSF1, and ACTG2

[107]. The identified biomarkers are functionally connected, indicating coordinated molecular regulation in obesity associated metabolic dysfunctions, instead of gene-level alterations [46]. Pathway enrichment analysis also revealed high enrichment of oxidative phosphorylation, reactive oxygen species signaling, glycogen metabolism, insulin resistance and mitochondrial metabolic pathways [117,121].

Findings from predictive modelling combined with explainability analysis and systems biology validation add to the confidence of biomarkers identified and underpin the potential for obesity biomarker discovery and precision metabolic medicine[44,98].

Overall, the findings highlight the potential of EAI and transcriptomic analysis coupled with systems biology for generating a biologically interpretable framework to identify molecular signatures of obesity and decipher complex mechanisms of metabolic disease[45].

CHAPTER 4

DISCUSSION

4.1 Biological Interpretation of Obesity-Associated Transcriptomic Signatures

In the present study, a cross-tissue transcriptomics, ensemble machine learning, explainable artificial intelligence and systems biology validation-based computational pipeline was leveraged to uncover biologically relevant obesity-associated molecular signatures [40,44,46]. The study combined the transcriptomic data from both skeletal muscle (GSE22309) and adipose tissue (GSE103766) to identify the systemic molecular changes that lead to obesity instead of focusing only on tissue-specific dysregulations. Overall, these data indicate that mitochondrial dysfunction, oxidative stress, metabolic inflexibility and/or unbalanced regulation of transcription are major components of obesity-associated transcriptomic profiles, consistent with obesity being viewed as a multisystem disorder of metabolism [24,28].

Ensemble learning offers significant advantages over traditional statistical modeling approaches because obesity-associated transcriptomic data is inherently heterogeneous and has complex gene–gene relationships.

Traditional machine learning models often operate as black boxes, limiting insights into biology and confidence in clinical application of computational results [42]. Of the prioritised transcriptomic biomarkers, genes such as BDH1, MRPS31, RARS2, STOML2, AKAP10, UCP3, NUCB2, SRSF1 and ACTG2 showed significant predictive value [40]. Some of these genes are directly linked to the metabolic pathways of mitochondria and oxidative phosphorylation, highlighting the role of mitochondrial dysfunction in obesity pathogenesis [24,73]. The roles of the mitochondrial transcriptomic biomarkers identified in the present study

thus reinforce the previous evidence that impaired mitochondrial bioenergetics is a hallmark mechanism of obesity progression [28]. Decreased metabolic flexibility and insulin resistance associated with obesity may be linked to dys regulations of UCP3, which affects mitochondrial proton transport and mitochondrial oxidative efficiency [22,126]. Likewise, STOML2, which acts in the organization of the mitochondrial membrane, and MRPS31, which is a component of the mitochondrial ribosome, also highlight the significance of a dysregulated mitochondrial integrity in obesity [73].

Protein–protein interaction network analysis further improved biological interpretation through identification of NUCB2, SRSF1 and ACTG2 as central hub genes [107].

4.2 Comparison with Previous Obesity Transcriptomic Studies

The present results confirm the molecular evidence and also provide a rationale machine-learning-based approach to biomarker prioritization [44]. Skeletal muscle dysfunction is a major factor in insulin resistance of obesity, due to reduced glucose uptake and metabolic inflexibility [20]. In the current study, several transcriptomic biomarkers, such as UCP3, BDH1 and STOML2, which have been related to mitochondrial metabolism, showed strong contribution towards obesity classification [73]. The results indicate significant transcriptomic changes in the mitochondria function of the obesity-associated skeletal muscle tissues [24].

Most previous obesity transcriptomic studies have used single tissue analysis and differential gene expression for biomarker discovery [38] – a method that is not used in this study. While these approaches offer valuable biological insights, they often fail to capture coordinated and systemic transcriptomic dysregulation in metabolically coupled tissues [37]. However, obesity impacts the activity of several organs at the same time, so integrative cross-tissue transcriptomic analysis can yield a more holistic view of obesity-associated molecular mechanisms [122]. In

the present study, therefore, an important methodological drawback of previous obesity studies is overcome, combining transcriptomic profiles of skeletal muscle and adipose tissue [37].

Although machine learning is being widely employed in biomedical research, explainability has not been adequately studied in obesity transcriptomic studies [44]. The majority of machine learning studies on obesity focus on prediction scores while lacking mechanistic insight into the drivers of prediction results [42]. This is mainly due to the fact that the incorporation of SHAP-based explainability analysis in the present study allowed the prioritization of obesity-associated transcriptomic biomarkers with a biological interpretation [45], which constitutes an important contribution.

4.3 Biological implications of mitochondrial dysfunction in obesity

Mitochondrial dysfunction has emerged as a key mechanism in obesity-associated metabolic disease, due to its role in ATP production, fatty acid oxidation and reactive oxygen species homeostasis [25]. The high transcriptomic priority of the mitochondrial-related genes BDH1, UCP3, MRPS31 and STOML2 reinforces the hypothesis that obesity progression is characterized by a broad mitochondrial metabolic dysfunction [73]. A reduction in mitochondrial respiration impairs energy utilization and is associated with a build-up of intracellular lipid intermediates which increases metabolic inflexibility and insulin resistance [22].

Chronic low-level inflammation is then caused due to oxidative stress, which leads to endothelial dysfunction and development of metabolic disorders associated with obesity [29,70]. Thus, transcriptomic markers related to mitochondrial dysfunction could be clinically relevant molecular markers of metabolic dysfunction and obesity onset [24].

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

Obesity is a multifactorial metabolic disorder characterized by coordinated dysregulations of adipose tissue function, skeletal muscle metabolism, mitochondrial bioenergetics, oxidative stress and inflammatory signaling pathways [10,24,29]. To this end, the present work sought to establish an integrated computational framework that leverages cross-tissue transcriptomics, ensemble machine learning, explainable artificial intelligence and systems biology validation to uncover obesity-associated transcriptomic biomarkers and boost the understanding of obesity development mechanisms [40,44,46].

To examine molecular changes in metabolically interconnected tissues that are associated with obesity, the present investigation utilized publicly available transcriptomic data from two types of tissue: skeletal muscle (GSE22309) and adipose tissue (GSE103766) [35,37]. When using transcriptomic profiles from two tissues, their integration allowed to identify systemic molecular signatures, which were not restricted to tissue-specific dysregulation [122]. The coordinated dysregulation of multiple organs involved in glucose uptake, lipid metabolism, and endocrine homeostasis of obesity makes cross-tissue transcriptomic integration significantly enhance the biological comprehensiveness and biomarker discovery potential [20,37].

In this regard, the implementation of machine learning showed that the ensemble learning methods outperformed the conventional statistical and classification models, especially the Random Forest and the XGBoost method [41]. The optimized Random Forest (RF) classifier

was among the evaluated algorithms that achieved equal levels of predictive performance and good discrimination between obesity-associated and control transcriptomic phenotypes [90,91]. The classification ability seen here demonstrates the success of ensemble learning methods in extracting biologically-relevant transcriptomic patterns from large datasets of obesity [89].

A significant gain in the interpretability of machine learning predictions arose from the integration of the explainable artificial intelligence and the capacity to prioritize obesity-associated transcriptomic biomarkers in a biologically relevant manner [44]. The genes BDH1, MRPS31, RARS2, STOML2, AKAP10, UCP3, NUCB2, SRSF1 and ACTG2 showed statistically significant associations with obesity classification outcomes [40]. The strong association of these biomarkers with mitochondrial metabolism, oxidative phosphorylation, transcriptomic regulation and cytoskeletal organization, supported the hypothesis that obesity progression is accompanied by broad metabolic and molecular dysfunction [24,36].

Protein–protein interaction network analysis also revealed that NUCB2, SRSF1 and ACTG2 are highly connected hub genes which may have a coordinated regulation role in obesity-related molecular pathways. Biological relevance of identified transcriptomic biomarkers was further confirmed by functional pathway enrichment analysis which revealed significant roles of pathways related to oxidative phosphorylation, reactive oxygen species signaling, glycogen metabolism, insulin resistance and mitochondrial dysfunction [117,121].

Taken together, all the results gained in the present study strongly suggest that mitochondrial dysfunction, oxidative stress, metabolic inflexibility and transcriptomic dysregulation are core molecular mechanisms involved in obesity progression [24,26]. The integrated computational framework was successfully demonstrated to be a powerful and biologically interpretable approach for discovering obesity-associated biomarkers using explainable machine learning, transcriptomics and systems biology [45,46].

This study therefore provides a contribution to the existing knowledge on the pathophysiology of obesity and points to the translational potential of EAI in precision metabolic medicine [44]. Biologically meaningful transcriptomic biomarkers can help for early diagnosis, molecular stratification and targeted therapeutic interventions to better manage obesity and its associated metabolic complications [98].

5.2 Future Scope

Experimental testing of prioritized transcriptomic biomarkers, including those related to mitochondrial metabolism and metabolic regulation [73,109] should be explored in future studies. Furthermore, the use of advanced explainable AI and deep learning approaches can improve predictive ability [44,98].

Future personalized diagnosis, machine learning and systems biology-based integrated transcriptomics approach can thus play a significant role in precision metabolic therapeutics and obesity diagnosis biomarker guided therapy[45,46].

An additional line of research that is being followed is the development of precision medicine frameworks to manage obesity [45]. The molecular obesity subgroups can be identified by the transcriptomic biomarkers discovered in the present study, which can help to guide personalized treatments [98]. Individual interventions for obesity that included mitochondrial dysfunction and/or oxidative stress or inflammatory signaling pathways may thus enhance therapeutic success and disease prevention [74].

Furthermore, larger and ethnically diverse patient cohorts might be included in future transcriptomic studies to increase biomarker replications and increase the applicability of the findings to patients in clinical settings [38]. Obesity is a highly variable phenotype, with genetic,

environmental, and lifestyle factors contributing to the differences, so larger population-based studies are needed to confirm transcriptomic signatures across a wider range of populations [11]. In summary, future studies using multi-omics and explainable AI and systems biology could significantly contribute to the understanding of obesity-related metabolic dysfunction and assist in the design of precision metabolic medicine [45,46].

REFERENCES

- [1] Obesity: Preventing and Managing the Global Epidemic. Geneva: World Health Organization; 2000.
- [2] Bray GA, Kim KK, Wilding JPH, et al. Obesity: a chronic relapsing progressive disease process. World Obesity Federation. 2017;18(7):715–723.
- [3] Hill JO, Wyatt HR, Peters JC. Energy balance and obesity. *Circulation*. 2012;126(1):126–132.
- [4] Hruby A, Hu FB. The epidemiology of obesity: a big picture. *Pharmacoeconomics*. 2015;33(7):673–689.
- [5] World Health Organization. Obesity and overweight. Geneva: World Health Organization; 2021.
- [6] Haslam DW, James WPT. Obesity. *The Lancet*. 2005;366(9492):1197–1209.
- [7] Kopelman PG. Obesity as a medical problem. *Nature*. 2000;404(6778):635–643.
- [8] Blüher M. Obesity: global epidemiology and pathogenesis. *Nature Reviews Endocrinology*. 2019;15(5):288–298.
- [9] Després JP. Body fat distribution and risk of cardiovascular disease: an update. *Circulation*. 2012;126(10):1301–1313.
- [10] Saltiel AR, Kahn CR. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature*. 2001;414(6865):799–806.
- [11] Friedman JM. Leptin and the endocrine control of energy balance. *Nature Metabolism*. 2019;1(8):754–764.
- [12] Morton GJ, Meek TH, Schwartz MW. Neurobiology of food intake in health and disease. *Nature Reviews Neuroscience*. 2014;15(6):367–378.
- [13] Klok MD, Jakobsdottir S, Drent ML. The role of leptin and ghrelin in the regulation of food intake and body weight in humans. *Obesity Reviews*. 2007;8(1):21–34.
- [14] Myers MG Jr, Leibel RL, Seeley RJ, Schwartz MW. Obesity and leptin resistance: distinguishing cause from effect. *Trends in Endocrinology & Metabolism*. 2010;21(11):643–651.
- [15] Izquierdo AG, Crujeiras AB, Casanueva FF, Carreira MC. Leptin, obesity, and leptin resistance: where are we 25 years later? *Nutrients*. 2019;11(11):2704.
- [16] Kershaw EE, Flier JS. Adipose tissue as an endocrine organ. *The Journal of Clinical Endocrinology & Metabolism*. 2004;89(6):2548–2556.
- [17] Ouchi N, Parker JL, Lugus JJ, Walsh K. Adipokines in inflammation and metabolic disease. *Nature Reviews Immunology*. 2011;11(2):85–97.
- [18] Rosen ED, Spiegelman BM. What we talk about when we talk about fat. *Cell*. 2014;156(1–2):20–44.
- [19] Shoelson SE, Herrero L, Naaz A. Obesity, inflammation, and insulin resistance. *Gastroenterology*. 2007;132(6):2169–2180.
- [20] DeFronzo RA, Tripathy D. Skeletal muscle insulin resistance is the primary defect in type 2 diabetes. *Diabetes Care*. 2009;32(Suppl 2):S157–S163.
- [21] Thibaud D, DeFronzo RA, et al. The effect of graded doses of insulin on total glucose uptake, glucose oxidation and glycogen synthesis in man. *Diabetes*. 1982;31(11):957–963.
- [22] Kelley DE, He J, Menshikova EV, Ritov VB. Dysfunction of mitochondria in human skeletal muscle in type 2 diabetes. *Diabetes*. 2002;51(10):2944–2950.

- [23] Samuel VT, Shulman GI. Mechanisms for insulin resistance: common threads and missing links. *Cell*. 2012;148(5):852–871.
- [24] Bournat JC, Brown CW. Mitochondrial dysfunction in obesity. *Current Opinion in Endocrinology, Diabetes and Obesity*. 2010;17(5):446–452.
- [25] Spinelli JB, Haigis MC. The multifaceted contributions of mitochondria to cellular metabolism. *Nature Cell Biology*. 2018;20(7):745–754.
- [26] Lowell BB, Shulman GI. Mitochondrial dysfunction and type 2 diabetes. *Science*. 2005;307(5708):384–387.
- [27] Furukawa S, Fujita T, et al. Increased oxidative stress in obesity and its impact on metabolic syndrome. *The Journal of Clinical Investigation*. 2004;114(12):1752–1761.
- [28] Montgomery MK, Turner N. Mitochondrial dysfunction and insulin resistance: an update. *Endocrine Connections*. 2015;4(1):R1–R15.
- [29] Gregor MF, Hotamisligil GS. Inflammatory mechanisms in obesity. *Annual Review of Immunology*. 2011;29:415–445.
- [30] Weisberg SP, McCann D, et al. Obesity is associated with macrophage accumulation in adipose tissue. *The Journal of Clinical Investigation*. 2003;112(12):1796–1808.
- [31] Hotamisligil GS. Inflammation and metabolic disorders. *Nature*. 2006;444(7121):860–867.
- [32] Van Gaal LF, Mertens IL, De Block CE. Mechanisms linking obesity with cardiovascular disease. *Nature*. 2006;444(7121):875–880.
- [33] Principles of Gene Manipulation and Genomics. 8th ed. Oxford: Blackwell Publishing; 2006.
- [34] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009;10(1):57–63.
- [35] Barrett T, Wilhite SE, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*. 2013;41(Database issue):D991–D995.
- [36] Keller P, Gburcik V, et al. Gene-chip studies of adipogenesis-regulated genes in obesity and insulin resistance. *International Journal of Obesity*. 2011;35(4):462–472.
- [37] Lotta LA, Scott RA. Multi-tissue transcriptomic integration in metabolic disorders. *Nature Genetics*. 2015;47(8):949–956.
- [38] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*. 2015;16(6):321–332.
- [39] Tarca AL, Romero R, et al. Machine learning and its applications to transcriptomics. *BMC Bioinformatics*. 2007;8(Suppl 7):S3.
- [40] Chen R, Snyder M. Promise of personalized omics and machine learning in medicine. *Nature Reviews Genetics*. 2013;14(4):247–257.
- [41] Libbrecht MW, Noble WS. Machine learning in genetics and genomics. *Nature Reviews Genetics*. 2015;16(6):321–332.
- [42] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019;1(5):206–215.
- [43] Samek W, Wiegand T, Müller KR. Explainable Artificial Intelligence: understanding, visualizing and interpreting deep learning models. *ITU Journal*. 2017;1(1):39–48.
- [44] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*. 2017;30:4765–4774.
- [45] Holzinger A, Langs G, et al. What do we need to build explainable AI systems for the medical domain? *arXiv preprint*. 2017;arXiv:1712.09923.
- [46] Ideker T, Sharan R. Protein networks in disease. *Genome Research*. 2008;18(4):644–652.

- [47] Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*. 2012;8(2):e1002375.
- [48] Szklarczyk D, et al. STRING v11: protein–protein association networks with increased coverage. *Nucleic Acids Research*. 2019;47(D1):D607–D613.
- [49] Cannon B, Nedergaard J. Brown adipose tissue: function and physiological significance. *Physiological Reviews*. 2004;84(1):277–359.
- [50] Virtanen KA, et al. Functional brown adipose tissue in healthy adults. *New England Journal of Medicine*. 2009;360(15):1518–1525.
- [51] Kadowaki T, Yamauchi T. Adiponectin and adiponectin receptors. *Endocrine Reviews*. 2005;26(3):439–451.
- [52] Arita Y, Kihara S, et al. Paradoxical decrease of an adipose-specific protein, adiponectin, in obesity. *Biochemical and Biophysical Research Communications*. 1999;257(1):79–83.
- [53] Yamauchi T, Kadowaki T. Adiponectin receptor as key player in healthy longevity and obesity-related diseases. *Cell Metabolism*. 2013;17(2):185–196.
- [54] Ye J. Emerging role of adipose tissue hypoxia in obesity and insulin resistance. *International Journal of Obesity*. 2009;33(1):54–66.
- [55] Sun K, Kusminski CM, Scherer PE. Adipose tissue remodeling and obesity. *The Journal of Clinical Investigation*. 2011;121(6):2094–2101.
- [56] Lumeng CN, Saltiel AR. Inflammatory links between obesity and metabolic disease. *The Journal of Clinical Investigation*. 2011;121(6):2111–2117.
- [57] Hirosumi J, et al. A central role for JNK in obesity and insulin resistance. *Nature*. 2002;420(6913):333–336.
- [58] Tilg H, Moschen AR. Adipocytokines: mediators linking adipose tissue, inflammation and immunity. *Nature Reviews Immunology*. 2006;6(10):772–783.
- [59] Divoux A, et al. Fibrosis in human adipose tissue: composition, distribution and link with metabolic dysfunction. *Diabetes*. 2010;59(11):2817–2825.
- [60] Sun K, Tordjman J, Clément K, Scherer PE. Fibrosis and adipose tissue dysfunction. *Cell Metabolism*. 2013;18(4):470–477.
- [61] Ibrahim MM. Subcutaneous and visceral adipose tissue: structural and functional differences. *Obesity Reviews*. 2010;11(1):11–18.
- [62] Wajchenberg BL. Subcutaneous and visceral adipose tissue: their relation to metabolic syndrome. *Endocrine Reviews*. 2000;21(6):697–738.
- [63] Hardy OT, Czech MP, Corvera S. What causes the insulin resistance underlying obesity? *Current Opinion in Endocrinology, Diabetes and Obesity*. 2012;19(2):81–87.
- [64] Ritov VB, et al. Deficiency of subsarcolemmal mitochondria in obesity and type 2 diabetes. *Diabetes*. 2005;54(1):8–14.
- [65] Puigserver P, Spiegelman BM. Peroxisome proliferator-activated receptor gamma coactivator 1 alpha (PGC-1 α): transcriptional coactivator and metabolic regulator. *Endocrine Reviews*. 2003;24(1):78–90.
- [66] Handschin C, Spiegelman BM. The role of exercise and PGC1 α in inflammation and chronic disease. *Nature*. 2008;454(7203):463–469.
- [67] Sena LA, Chandel NS. Physiological roles of mitochondrial reactive oxygen species. *Molecular Cell*. 2012;48(2):158–167.
- [68] Tiganis T. Reactive oxygen species and insulin resistance: the good, the bad and the ugly. *Trends in Pharmacological Sciences*. 2011;32(2):82–89.
- [69] Maddux BA, et al. Protection against oxidative stress-induced insulin resistance in rat models. *American Journal of Physiology*. 2001;281(2):E412–E420.

- [70] Vincent HK, Taylor AG. Biomarkers and potential mechanisms of obesity-induced oxidant stress in humans. *International Journal of Obesity*. 2006;30(3):400–418.
- [71] Keaney JF Jr, et al. Obesity and systemic oxidative stress: clinical correlates of oxidative biomarkers. *Arteriosclerosis, Thrombosis, and Vascular Biology*. 2003;23(3):434–439.
- [72] Mootha VK, et al. PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*. 2003;34(3):267–273.
- [73] Schrauwen P, Hesselink MKC. Oxidative capacity, lipotoxicity and mitochondrial dysfunction in obesity. *Physiology & Behavior*. 2004;83(4):599–605.
- [74] Crunkhorn S, et al. Modulation of mitochondrial metabolism as therapeutic intervention in obesity. *Nature Reviews Drug Discovery*. 2007;6(5):435–446.
- [75] Schena M, et al. Quantitative monitoring of gene expression patterns with complementary DNA microarray. *Science*. 1995;270(5235):467–470.
- [76] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*. 2008;5(7):621–628.
- [77] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*. 2002;30(1):207–210.
- [78] Pattern Recognition and Machine Learning. Bishop CM. New York: Springer; 2006.
- [79] An Introduction to Statistical Learning. James G, Witten D, Hastie T, Tibshirani R. New York: Springer; 2013.
- [80] Libbrecht MW, Noble WS. Machine learning in genetics and genomics. *Nature Reviews Genetics*. 2015;16(6):321–332.
- [81] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*. 2015;13:8–17.
- [82] Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*. 1967;13(1):21–27.
- [83] Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. Wiley; 2013.
- [84] Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995;20(3):273–297.
- [85] Rish I. An empirical study of the naïve Bayes classifier. *IJCAI Workshop on Empirical Methods in Artificial Intelligence*. 2001;3(22):41–46.
- [86] Quinlan JR. Induction of decision trees. *Machine Learning*. 1986;1(1):81–106.
- [87] Breiman L, Friedman J, Olshen R, Stone C. *Classification and Regression Trees*. Belmont: Wadsworth; 1984.
- [88] Hawkins DM. The problem of overfitting. *Journal of Chemical Information and Computer Sciences*. 2004;44(1):1–12.
- [89] Dietterich TG. Ensemble methods in machine learning. *Multiple Classifier Systems*. 2000;1857:1–15.
- [90] Breiman L. Random forests. *Machine Learning*. 2001;45(1):5–32.
- [91] Biau G, Scornet E. A random forest guided tour. *TEST*. 2016;25(2):197–227.
- [92] Chen X, Ishwaran H. Random forests for genomic data analysis. *Genomics*. 2012;99(6):323–329.
- [93] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785–794.
- [94] Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*. 2001;29(5):1189–1232.

- [95] Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*. 2013;7:21.
- [96] Lundberg SM, et al. Explainable AI for trees: from local explanations to global understanding. *Nature Machine Intelligence*. 2020;2(1):56–67.
- [97] Guyon I, Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research*. 2003;3:1157–1182.
- [98] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*. 2019;25(1):44–56.
- [99] Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*. 2007;3(9):1724–1735.
- [100] Lundberg SM, Erion G, et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*. 2020;2(1):56–67.
- [101] Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. Munich: Leanpub; 2022.
- [102] Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*. 2018;2(10):749–760.
- [103] Merrick L, Taly A. The explanation game: explaining machine learning models using SHAP values. *arXiv preprint*. 2020;arXiv:2006.16234.
- [104] Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*. 2004;5(2):101–113.
- [105] Vidal M, Cusick ME, Barabási AL. Interactome networks and human disease. *Cell*. 2011;144(6):986–998.
- [106] Menni C, et al. Systems biology approaches for obesity biomarker discovery. *Genome Medicine*. 2017;9(1):1–14.
- [107] Jeong H, Mason SP, et al. Lethality and centrality in protein networks. *Nature*. 2001;411(6833):41–42.
- [108] Muoio DM, Newgard CB. Molecular and metabolic mechanisms of insulin resistance and β -cell failure in type 2 diabetes. *Nature Reviews Molecular Cell Biology*. 2008;9(3):193–205.
- [109] Azuaje F. Computational models for predicting disease genes and therapeutic targets. *Briefings in Bioinformatics*. 2017;18(5):820–836.
- [110] Szklarczyk D, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*. 2019;47(D1):D607–D613.
- [111] Jensen LJ, et al. STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*. 2009;37(Database issue):D412–D416.
- [112] Szklarczyk D, et al. The STRING database in 2021: customizable protein–protein networks and functional characterization of user-uploaded gene sets. *Nucleic Acids Research*. 2021;49(D1):D605–D612.
- [113] Shannon P, Markiel A, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*. 2003;13(11):2498–2504.
- [114] Assenov Y, Ramírez F, et al. Computing topological parameters of biological networks. *Bioinformatics*. 2008;24(2):282–284.
- [115] Tang Y, et al. CytoNCA: a Cytoscape plugin for centrality analysis and evaluation of protein interaction networks. *Biosystems*. 2015;127:67–72.
- [116] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. 2000;28(1):27–30.

- [117] Kanehisa M, Furumichi M, et al. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*. 2023;51(D1):D587–D592.
- [118] Jassal B, et al. The Reactome pathway knowledgebase. *Nucleic Acids Research*. 2020;48(D1):D498–D503.
- [119] Kuleshov MV, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*. 2016;44(W1):W90–W97.
- [120] Chen EY, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 2013;14:128.
- [121] Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences USA*. 2005;102(43):15545–15550.
- [122] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biology*. 2017;18(1):83.
- [123] Anczuków O, Krainer AR. Splicing-factor alterations in cancer and metabolic disease. *Nature Reviews Cancer*. 2016;16(6):413–430.
- [124] Oh IS, et al. Identification of nesfatin-1 as a satiety molecule in the hypothalamus. *Nature*. 2006;443(7112):709–712.
- [125] Rønnev-Jessen L, Petersen OW. A function for filamentous actin in cell motility and tissue organization. *Journal of Cell Biology*. 1996;134(1):67–80.

LIST OF CONFERENCES

1. INTERNATIONAL CONFERENCE ON: “BioFrontiers 2026: Trends, Technologies & Translational Research” hosted by the Department of Microbiology SRICT-ISR, UPL University of Sustainable Technology (UUST) on 6th and 7th March 2026



CERTIFICATE OF ACHIEVEMENT

This is to certify that KANAN has secured 1st Rank in the UG/PG Student Oral Presentation category for the presentation entitled A Computational Framework for Identifying Obesity Associated Gene Signatures at the 1st International Conference “BioFrontiers 2026: Trends, Technologies & Translational Research”, held on 6–7 March 2026.

The conference was organized by the Department of Microbiology, SRICT – Institute of Science and Research, UPL University of Sustainable Technology (UUST), Ankleshwar, Gujarat, India, in association with Savli Technology & Business Incubator (STBI), Savli–Baroda.

This recognition is awarded for outstanding research presentation and scientific contribution.

Dr. Prashant Arya
Coordinator
Assistant Professor
SRICT-ISR, UUST

Dr. Trupti Patel
Asso. Prof. & Head
Department of Microbiology
SRICT-ISR, UUST

Dr. Snehal Lokahandwala
Dean Science and Sustainability,
SRICT-ISR, UUST

Dr. Shrikant J. Wagh
Provost, UUST

2. "3rd International Conference on Advances in Engineering and Medical Sciences 2026" organised " held during 17th & 18th April 2026 at International School of Technology and Sciences for Women, India.

3rd INTERNATIONAL CONFERENCE ON ADVANCES IN ENGINEERING AND MEDICAL SCIENCES 2026

ORGANISED BY

INTERNATIONAL SCHOOL OF TECHNOLOGY AND SCIENCES FOR WOMEN (AUTONOMOUS)
NH-16, EAST GONAGUDEM, RAJANAGARAM, EAST GODAVARI-533297 | WWW.ISTS.AC.IN

IN COLLABORATION WITH

MANIPAL UNIVERSITY COLLEGE, MALAYSIA
SAMARKAND STATE UNIVERSITY, UZBEKISTAN
UNIVERSITY OF TECHNOLOGY AND APPLIED SCIENCES, IBRI SULTANATE OF OMAN

IN ASSOCIATION WITH

ORGANIZATION OF SCIENCE AND INNOVATIVE ENGINEERING & TECHNOLOGY (OSIET)

Certificate of Presentation

This is to certify that *Mr/Mrs/Dr* **Kanan**
from **Delhi Technological University, Delhi** *has presented a paper title*
**A Computational Framework for Identifying Obesity-Associated Gene Signatures Using Human Adipose
and Skeletal Muscle Transcriptomes**
.....
*in the "3rd International Conference on Advances in Engineering and Medical Sciences" held on
17th & 18th April 2026 at International School Of Technology And Sciences For Women, India.*

Accredited by








Dr. Y. Rajasree Rao
Principal & Conference Chair





K. Janani
CEO





Dr. Antony V. Samrot
Director (Research, Innovation and Postgraduate Studies)
Manipal University College, Malaysia





Dr. Christo Ananth
Professor





Dr. Akhatov Akmal Rustamovich
Vice Rector (International Cooperation)



Approved by








DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

PLAGARISM VERIFICATION

Title of the Thesis: A COMPUTATIONAL FRAMEWORK FOR IDENTIFYING GENE SIGNATURES USING HUMAN ADIPOSE AND SKELETAL MUSCLES TRANSCRIPTOMES

Total Pages: 75

Name of the Student: KANAN (2K24/BIO/14)

Supervisor: Prof. ASMITA DAS

Department of Biotechnology, Delhi Technological University, Delhi-110042

This is to report that the above thesis was scanned for similarity detection. Process and outcome are given below:

Software Used: Turnitin, Similarity Index: 4%, Total Word Count: 15396

DATE: 27/05/26

Kanan

**CANDIDATE'S
SIGNATURE**

ASMITA DAS

**PROF. ASMITA DAS
SUPERVISOR**