Comparative Analysis of Efficient Deep Learning Models for Breast Cancer Identification Using Relevant Genes

by

Student Name: Md Abu Saad
Student ID: 20301173
Student Name: B. M Anjum Ul Muqset
Student ID: 20301223
Student Name: Chowdhury Rifat Ahmad Shopnil
Student ID:20301167
Student Name: Md.Mostafizur Rahman
Student ID: 20301163

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md Abu Saad	B. M Anjum Ul Muqset
20301173	20301223
Chowdhury Rifat Ahmad Shopnil	Md.Mostafizur Rahman
20301167	20301163

Approval

1. Md Abu Saad (20301173)
2. B. M Anjum Ul Muqset (20301223)
3. Chowdhury Rifat Ahmad Shopnil (20301167)
4. Md.Mostafizur Rahman (20301163)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on September, 2024.

Examining Committee:
Supervisor:
(Member)

Md Ashraful Alam, PhD

Associate Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD

Associate Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi

Associate Professor and Chairperson
Department of Computer Science and Engineering
BRAC University

Abstract

Cancer remains a formidable global health challenge, with early detection critical in improving patient outcomes. In this context, applying deep learning techniques to relevant gene analysis has emerged as a promising avenue for enhancing cancer detection and diagnosis. This study presents an investigation into utilizing dimensionality reduction methods and deep learning techniques to analyze gene sequences with the primary aim of detecting breast cancer. Various dimensionality reduction techniques reduce the amplitude by selecting a subset of relevant characteristics or variables from a larger collection of available features. The selected relevant features are then transformed into meaningful representations obtained using SDAE (Stacked Denoising Autoencoder) which are used to familiarize the data with different noise and outliers. Various expert learning architectures have been studied to evaluate the effectiveness of compact functions resulting from SDAE transforms. In addition, we apply discrete classification algorithms such as SVM, ANN, and SVM-RBF to distinguish between cancer and normal samples. The primary objective is cancer detection, encompassing the identification of breast cancer in its early stages, the recurrence of cancer, and the pathological response based on the genetic data. The study contributes to the growing body of knowledge in bioinformatics and medical research. It holds the potential to translate its findings into practical clinical applications, thereby advancing our ability to combat this devastating disease. This work represents a significant step towards realizing more effective and precise cancer diagnostics, offering hope for improved patient care and outcomes in the fight against cancer.

Keywords: Cancer Identification; Deep Learning; Dimensionality Reduction; Feature learning; Machine Learning; Stacked Denoising Autoencoder; Gene sequences; Evaluation metrics

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our supervisor Dr. Md Ashraful Alam for his kind support and advice in our work. He helped us whenever we needed help.
And finally to our parents without their support, it may not be possible. With their kind support and prayer, we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ANN Artificial Neural Network

CNN Convolutional Neural Network

DAE Denoising Auto Encoder

FS Feature Selection

$KPCA$ Kernel Principle Component Analysis

PCA Principle Component Analysis

RP Random Projection

$SDAE$ Stacked Denoising Auto Encoder

SVM Support Vector Machine

$SVM - RBF$ Support Vector Machine Kernel

$TCGA$ The Cancer Genome Atlas Program

Chapter 1

Introduction

Cancer is one of the most deadly diseases suffered by mankind. It is defined by the uncontrolled growth and spread of cells in the body. These unnatural cells can form tumors, invade, and spread to nearby tissues, and in some cases, to other body parts as well. Cancer is caused because of genetic mutation, environmental factors, and lifestyle as well and it can occur in almost any tissue or organ of a person's body. There are many varied types of this disease, each with its own characteristics. The identification of cancer has been in research since the beginning of the advancement of medical history. However, the complex nature and numerous types have proved extremely challenging to understand and diagnose this unwelcome disease.

The early detection of cancer plays a vital part in cancer diagnosis which improves survival rates in the long run. For this early detection, medical imaging is an important technique and it has been well known and widely employed to detect, monitor, and chase up even after treatments [5]. However, manually implementing a massive number of medical images can be a tough and prolonged task and can cause human mistakes. For this reason, CAD(computer-aided diagnosis) systems were set in motion in the 1980s to improve the efficiency of the images and help interpret them [4].

Due to the severity of the cancer outcome, timely and precise diagnosis is a critical factor in enhancing treatment efficacy. Although there has been much research on the identification of cancer, gene identification proved to be a field of great curiosity among seekers, According to K. Kourou (2015) machine learning approaches have greatly increased the accuracy of the extracted data by 15%-20% in the previous decade[14]. Machine learning methods that are used to reduce the dimensionality of various data and proper extraction of that required data have proved to be effective but still, there are some limitations in the interpretation of the most significant data among the complex genome datasets. According to P. Danaee and R. Ghaeini (2017), The available data on a very large scale and the noise associated with them make such tasks challenging. Furthermore in [21], there occurs a "dimensionality curse as a consequence of the incompatibility between a large number of genes and a comparatively small number of data points. Among many machine learning approaches to solve such issues of reducing noise, SDAE proved to be more effective in recent times. SDAE is built using deep learning neural networks which possess a unique ability to extract and learn meaningful features from complex datasets, which

makes it ideal for use in the difficult extraction of precise and highly relevant genes and serves as a biomarker for cancer diagnosis. In this paper, we dedicate ourselves to exploring the boundless possibilities and potentials in the practical applications of the SDAE. Inquiring into the theoretical fundamentals of SDAE, we shall provide an in-depth methodology to show a practical implication of this deep learning method to show how helpful it would be to unravel the complexities of genetic information. In this study, we utilized stacked denoising autoencoders (SDAE) to convert high-dimensional, noisy data on gene expression to a lower-dimensional, unambiguous characterization.

However, it is critical to recognize that, while SDAEs provide revolutionary opportunities, they also provide complications and ethical concerns. We will thoroughly investigate these variables, including data privacy, model interpretability, and the possibility of bias in AI-driven diagnostics. We can ensure that the incorporation of SDAEs into cancer diagnosis is both conscientious and equitable by dealing with these challenges straight on.

1.1 Research Problems

Cancer is one of the top causes of mortality globally [15] and observation by the American Cancer Society (ACS) reveals that 1,958,310 new cancer patients are to appear and 609,820 people are expected to die from cancers in the USA in 2023 [33]. According to the survey, even if the mortality rates of cancer have declined in the past few years with the new and advanced treatment techniques, as breast, prostate, and uterine corpus cancer occurrences are increasing, future progress might be attenuated. The increase in variations of the types of the disease and change in characteristics in the cancer biomarkers are to be a challenge for detection and diagnosis.

According to [3], [8], [12], one of the main reasons for failed treatments of cancer is because of the delayed start of treatment and that happens because of not being able to detect unnaturally growing cells in the tissue quicker. In many cases, the patient's cancer does not show understandable symptoms until the late stages and the medical professional fails to identify the disease properly. The delayed diagnosis of cancer can lead to extreme consequences as the patient may miss the critical window for the treatment to be most effective. Failing to do so gives cancer cells a chance to grow, affect other tissues, and reach other parts of the body which become harder to treat. Thus, in some of these cases, the cancer becomes incurable and leads to the patient's death.

Medical imaging is crucial to detect cancer early; since the arrival of CAD systems, the efficiency of interpreting medical images has improved a lot. In CAD systems, advanced machine learning techniques are employed for medical imaging. To adapt to these machine learning techniques, various feature extraction models were assessed for different image modalities and various cancer types to detect mass mammograms [15]. However, according to [24], the studies of adopting bilateral image subtraction, a difference of Gaussian, and Laplacian of Gaussian filters as feature extractors had a lot of weaknesses and these limit the further improvement

of the execution of CAD systems. Therefore, the significance of representative learning has been addressed. Deep learning is a representation learning technique that extracts hierarchical image representation from the images.

In [24], a survey is done on various deep learning models being used to detect different types of cancer and afterward, they came to derive some challenges. One of the biggest challenges was the lack of large training datasets even though there were millions of medical images, most were kept confidential. Moreover, some of the open-source data files contained raw image data and required extra effort to evaluate.

A frequent problem in medical imaging applications is the limited number of labeled training data available [26]. In [29], the challenges faced were scarce, sparse, noisy, and image-level annotations of images. For these weak annotations of images, extracting useful features becomes difficult, and hard to correctly identify cancer biomarkers. To confront this problem of limited labeled data, unsupervised or self-supervised algorithms are needed to be explored. In [26], the capability of unsupervised learning to determine malignancy was explored. As annotations were required for the radiologists in most medical imaging tasks and acquiring labels for the training of machine learning models is heavy and costly in comparison to other computer vision tasks, clustering was used to obtain the initial set of labels and afterward, they were filtered with SVM. This approach obtained better results than other methods of evaluation metrics.

Some problems arise about the robustness of the data while working with the SDAE model. According to Vincent et al [9], the SDAEs are generally built to be robust to input noise, but they may struggle in several conditions such as high noise levels, structured noise, and optimal noise levels. When the noise level in data emerges too high SDAEs are found struggling to iterate properly. Excessive noise can be a big challenge for the autoencoder to extract meaningful representations. When it comes to the question of structured noise, it possesses another struggling point because SDAEs are mostly effective on random noise. Choosing the optimal noise levels to train the SDAE is very crucial, too much noise can obstruct the learning process and inadequate noise can limit the model's capabilities.

Gene expression microarrays provide a bunch of information on pattern expression and cancer pathways which can be utilized in the diagnosis and prognosis of cancer. Furthermore, they help to discover new cancer subtypes and identify cancer signaling molecular markers and their complexity. However, according to [6] comprehensive analysis of microarray gene expression profiles is required which consists of high dimensional data spaces Although microarray data analysis falls deftly within pattern recognition and statistical learning, in no way the tasks are conventional as there is a lack of samples per dimension. Thus, a way of lowering dimensionality and getting a meaningful representation of the gene expression data is required to correctly identify cancer biomarkers .

Therefore, the question that this study is trying to answer is:

"Can our study compare the effectiveness of various deep learning models using gene expression data to identify the most accurate and reliable model for breast cancer detection? "

This research will answer the above question by exploring the SDAE model and its proficient use in extracting important features to detect cancer.

1.2 Research objectives

Our study primarily focuses on the implementation of integrating Feature selection with a Stacked denoising Autoencoder for dimensionality and noise-reduced genes for diagnosing breast cancer. The proposed model is then compared to certain well-known pre-trained models, such as PCA and KPCA, and the accuracy, loss, and other data are compared to show whether or not the suggested model meets the standards. Furthermore, this study aims to acquaint ourselves with the nuances of every model and examine them to contrast their outcomes and pinpoint detecting mistakes to take advantage of the vulnerabilities. This will also assist us in resolving those problems and producing a faster and more precise version of our suggested model.

To do that, we want to create an open-source, complex algorithm that combines deep learning with picture categorization. This method has the potential to revolutionize deep learning in the future. Our goal is to develop a model that is easier to use in any system and has fewer parameters and a lower degree of complexity.

The objectives of this research are:

1. Explore and develop an integrated methodology that combines effective dimensionality reduction algorithms with SDAE to enhance the extraction and representation of relevant features.

2.Adapt and design SDAE architectures capable of effectively denoising and capturing complex patterns in genomic and transcriptomic data, allowing for the extraction of meaningful representations related to breast cancer biology.

3. Investigate the ability of the combined dimensionality reduction + SDAE approach to reduce the dimensionality of the high-dimensional genomic data while preserving critical information and capturing subtle variations indicative of breast cancer subtypes.

4. Compare the performance of the optimal model against baseline models that use the techniques individually alone, as well as traditional machine learning models without dimensionality reduction techniques, to highlight the advantages of the proposed integrated approach.

5. Explore the clinical relevance and translational potential of the identified genomic

features, seeking to contribute insights that may inform personalized therapeutic strategies and enhance the understanding of the molecular landscape of breast cancer

Chapter 2

Literature Review

Machine learning and image-based learning approaches have been used for some time in the identification and diagnosis of cancer. Recent publications have discussed the use of deep learning algorithms to identify cancer biomarkers in many studies. This section aims to critically evaluate prior relevant work in the area of deep learning-based cancer diagnosis. Despite the difficulties of collecting reliable data from huge complicated databases, we analysed many ways.

A study done on deep learning and image-based cancer detection in [24] shows surveys done on multiple papers regarding the topic. In [16], created a model for feature extraction which follows the deep learning approach to identify mitosis in breast histopathological images. The suggested technique implemented a CNN model to extract the features then they were used to train a support vector machine (SVM) for mitosis detection. In a paper published in 2016, Spanhol et al. utilized AlexNet to build a CNN model to classify malignant tumors in breast histopathology pictures [17]. For classifying nuclei in breast cancer histology, Xu et al [13] indicated a stacked sparse auto-encoder (SSAE)-based approach. The SSAE was optimized through the training process using a greedy technique in which one hidden layer was taught at a time and the above layer's output was used for the training of the next hidden layer. Aside from using deep learning to detect tumors in histopathological images, other studies have focused on using deep learning models to identify breast cancer in mammographic images. In Wichakam et al [19]suggested a technique for mass detection on digital mammograms that combines deep CNN and SVM. On mammographic patches, a CNN model was constructed, and the output from the final fully connected layer was used as the image's high-level feature representation which trained an SVM for classification. Suzuki et al. suggested a transfer learning technique to train CNN models for mass identification in mammographic images due to a lack of training photos to train a deep CNN model[18]

In research work [31], the use of computer-aided systems in pathology labs is emphasized as it will complement human expertise and the early detection of cancer. The authors propose an automatic segmentation approach, followed by self-driven post-processing activities. This approach is based on Fourier Transform and it offers several advantages over existing techniques. This paper's[22] study offers an enhanced classification methodology combining feature selection, Genetic Algorithm (GA), and Convolutional Neural Network (CNN). In each iteration of the learning

process, the technique seeks to solve difficulties identified throughout the sample selection process. The use of denoising autoencoders (DAE) and stacked denoising autoencoders (SDAE) is studied as deep learning methods to derive useful features from genome data. The study is divided into two main tasks and it achieves the first task by getting high accuracy in discriminating thyroid cancer from healthy patients. However, the second task of extraction of comprehensible models was not completed as it was very challenging.

In [28], the authors developed a stacked denoising autoencoder model named MLP-SDAE to identify potential genes related to lung cancer. It used multilayer perceptron (MLP) for backpropagation and SDAE training for feature selection. MPL-SDAE was successfully categorized based on correlation coefficients and selected subsets of potential genes. This model outperformed other selection models such as SVM, SAM, BA, and GMM.Different machine-learning techniques were employed by Sterling Ramroach et al[27] to categorize cancer. They got a dataset for several cancer kinds from the COSMIC internet data site to aid in their study. They used a variety of models for machine learning, including the use of support vector machines, artificial neural networks, K closest neighbors, and random forest models. The researchers tested several primary cancer sites and cancer types. Due to its simple tuning and 100% classification accuracy, RF stood out above other algorithms by utilizing a multimodal ensemble strategy. In [23] an innovative approach to deep learning treatment of three distinct malignancies (LUAD, STAD, and BRCA) was offered. In this method, each classifier was trained independently using the supplied data to produce predictions, and these predictions were then used to guide a deep learning multimodal ensemble approach. The Knowledge Networks (KNNs), Statistical Machines (SVM), Decision Trees (DTs), Functional Randomization (RFs), and Gradient Boosting Decision Trees (GBDT) were the foundation of this method. Another research[2] shows that it is possible to use a feature selector based and various classifiers, which leads to reliable selection of the right gene in terms of classification accuracy as well as against more traditional methods.

Additionally,[25] classify tumor types based on deep learning. Because many tumors share the same biomarkers, they developed a way to identify a gene that can differentiate between different tumor types. They classified 33 tumor types using conventional neural networks (CNN) and using 2D large-scale RNA-seq data. They used a three-layer convolutional neural network to ensure good accuracy and speed. They scored 95.59 in the final. They also created special heat maps for each type using the idea of a controllable surveillance camera.

The aforementioned classification cannot handle genome-wide multidimensional data (such as metabolites, transcriptomes, and genomes). To address this issue, In research work [30] proposed a unified multi-task deep learning framework called OmiEmbed, which is built on the PyTorch deep learning library. This framework is used to collect high-dimensional omics data using a deep integration module and an upstream task module. The accuracy of the OmiEmbed-CNN method is more than 84.52% and the accuracy of the OmiEmbed-FC method is 87.54%. In addition, single-task accuracy is 96.76%, and multi-task accuracy is 97.33%.

Freitas et al [32] developed an ML model that can be used to determine cancer types based on certain microbial data. They trained Random Forest algorithms for five different types of cancer. The RF model achieved an accuracy of 72.35% to 96.04% depending on different cancer types. Combining patterns of gene expression and algorithms for supervised learning, it is achievable to identify potential cancer cells from healthy ones in a multitude of ways. According to Lee et al [7], feature learning methods and deep learning have a lot of applications in image and sound processing. These strategies have shown great promise in these fields to autonomously express the feature space using unlabeled data to improve the accuracy of subsequent classification problems. These features have been enhanced to enable learning in very high-dimensional feature spaces using new data attributes. Another approach that showed an interesting aspect is that of C. F. Alifers [1] where Alifers evaluated a small selection of gene expressions as a more limited set of characteristics using recursive feature reduction and univariate association filtering approaches. In their experiment, they used linear and polynomial kernel support vector machines and forward neural networks. The hypothesis of gene expression data combined with sophisticated learning algorithms makes a surprisingly effective diagnostic model for lung cancer types was proved by the outcome of their research. It was effective even with small samples and insufficient sample-to-function ratios.

Chapter 3

Background Information

3.1 Deep Learning

Deep learning is a type of machine learning in which artificial neural networks are used to model and interpret complex data patterns and representations. Deep learning algorithms are based on the structure and function of the human brain. They are composed of multiple layers of networked nodes, or neurons, which are organized into input, hidden, and output layers. Without explicit programming, these networks can learn abstract and hierarchical properties from unprocessed input, allowing them to perform tasks such as pattern recognition and prediction.

The depth of the neural network, or the several layers that allow it to perceive complicated relationships in data, is an important factor in deep learning. As each layer adds more abstract features, the system has a better understanding of sophisticated structures and representation. Deep learning is a valuable tool in a variety of industries, including technology, healthcare, and finance. It excels at tasks like audio and image recognition, natural language processing, and decision-making.

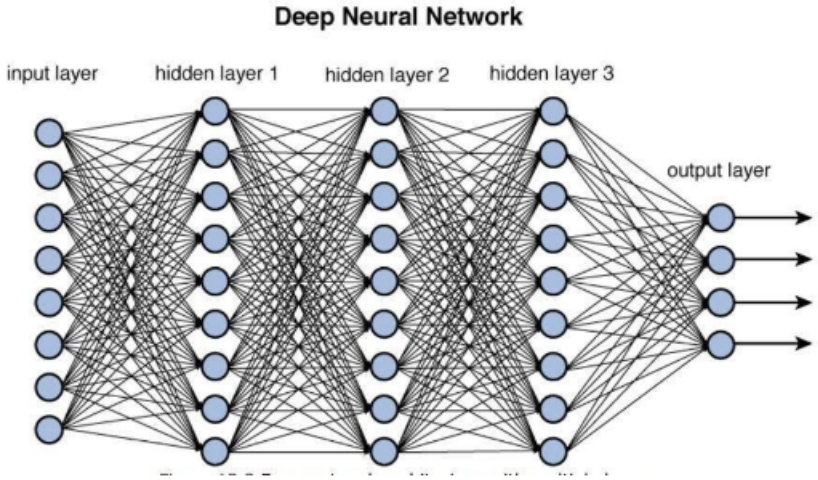

Figure 3.1: Deep Network Architecture with Multiple Layers

A deep learning model is trained by feeding it labeled data and using backpropagation to change the weights and biases of the connections between neurons. Through

this repeated learning process, the network is able to improve its performance on the given job and optimize its parameters. Deep learning has achieved tremendous success in recent years, surpassing traditional machine learning approaches in many applications and furthering artificial intelligence. Despite its triumphs, deep learning remains an ongoing research field, with challenges such as the need for large labeled datasets and computer resources

3.2 Feature Selection (FS)

One of the most critical phases in preparing data for machine learning models is feature selection. To improve the model's performance and efficiency, select a subset of relevant characteristics or variables from a larger collection of available features.

Consider a dataset with numerous features or properties that describe each instance. These attributes could include a variety of metrics, features, or properties associated with the data points. However, not every characteristic is equally important or improves a model's capacity to predict events. Including characteristics in the model that are redundant, superfluous, or even noisy may result in overfitting, increased computing costs, and poor interpretability. The purpose of feature selection is to identify and retain only the most informative information required for the model to make credible predictions. The technique comprises determining the relevance of each feature and its impact on the overall performance of the model.

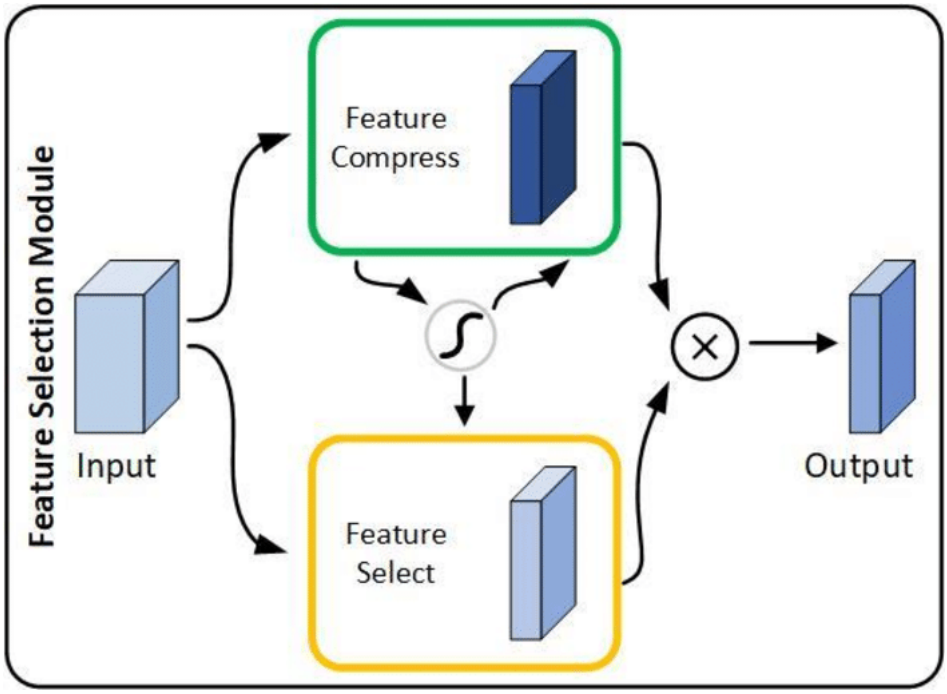

Figure 3.2: Structure Of Feature Selection

Practitioners can benefit from feature selection in a variety of ways. Removing irrelevant or superfluous characteristics might improve the model's accuracy and generalization to new, previously unknown data. Focusing on critical features reduces the

danger of overfitting, which occurs when a model performs well on training data but badly on new data. Furthermore, using a subset of features decreases computing costs, allowing the model to train faster and predict more efficiently. A model with fewer, more relevant elements is easier to grasp and comprehend, which is significant in applications requiring interpretability.

3.3 SDAE

The abbreviation SDAE stands for Stacked Denoising Autoencoder.Autoencoders are a type of artificial neural network designed for unsupervised learning. An autoencoder's primary function is to encode input data into a compressed representation, which it then decodes back to its original form. This method makes it easier to learn a more compact and effective representation of the input data.

Denoising autoencoder (DAE) is an alternative to conventional autoencoders. The basic idea behind a denoising autoencoder is to introduce deliberate noise into the input data and then train the network to reconstruct the original, clean input from the noisy version. Making the autoencoder deal with and eliminate noise helps it learn a reliable and practical representation of the data.

A stacked denoising autoencoder extends this concept by stacking multiple denoising autoencoder layers on top of one another. Each layer of the network is trained to extract higher-level features from the data, resulting in a hierarchical representation. While later layers can learn more complex and abstract features, the first layer may only learn basic characteristics such as edges and textures. To reduce the difference between the original input and the reconstructed clean input, the noisy input is fed through the network's stacked layers, and its weights are adjusted. The stacked architecture allows the network to automatically identify and encode relevant features in input data by learning hierarchical representations.

Data sets with errors, disturbances, or random or unpredictable variations that deviate from the true underlying patterns or information are referred to as noise data. Let's say you have a dataset that depicts a real-world phenomenon. However, for a variety of reasons, the data may not accurately represent the values you are attempting to quantify. When collecting data, one may encounter interference from external factors, measurement errors, or even inherent variability in the system under study. This interference adds noise to the dataset, making it less accurate and potentially more difficult to analyze.

Dealing with noisy data is a critical component of data analysis and machine learning. Environmental factors, human error during data entry, sensor inaccuracies, and even natural variations in the phenomenon under study can all contribute to noise. Unresolved noisy data can lead to inaccurate models, false insights, and lower machine-learning algorithm performance. This is where the Stacked Denoising Autoencoder (SDAE) comes into play. SDAE is a sophisticated system of interconnected layers, each of which plays a specific role in understanding and representing data. The primary innovation is its ability to handle noisy data gracefully. The first

layer of the SDAE is responsible for processing the raw input data, which includes noise. Rather than being overwhelmed by the noise, the network views it as a challenge. It intentionally introduces noise into the data, resulting in an altered input version. The network's goal is to filter out the additional noise while reconstructing the original, clean input from the noisy version. It gains the ability to distinguish between the distracting fluctuations introduced and the key characteristics of the data. As data passes through the layered architecture, each layer builds on the knowledge gained by the previous layer. Together, they improve the input's representation by gradually removing redundant noise and capturing deeper and meaningful qualities.

The network's goal is to recreate the original, clean input from the noisy version while filtering out any further noise. It gains the ability to discriminate between the fundamental elements of the data and the distracting variations that have been introduced. By the time the data progresses through the full stack, the SDAE does not only learn a hierarchical representation of the input, but it also masters handling and minimizing the effects of noisy elements. The network successfully trains itself to distinguish between signal and noise, allowing it to provide more robust and accurate representations of the underlying patterns in the data.In essence, the Stacked Denoising Autoencoder uses noisy input as a learning opportunity, teaching itself to extract valuable information even in the presence of disruptions.

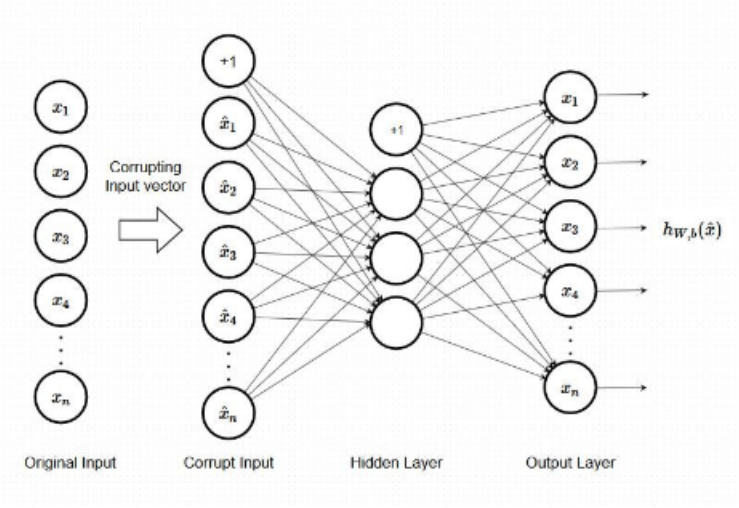

Figure 3.3: SDAE Architecture

In the case of high-dimensional datasets, each data point contains a large number of features. While this wealth of information is useful, it can also add complexity and make analysis difficult. In the case of high-dimensional datasets, each data point contains a large number of features. While this wealth of information is useful, it can also add complexity and make analysis difficult. As the SDAE processes your data through its stacked levels, each layer adds to a more refined representation of the input. This technique is more than just denoising; it also requires capturing the most important features of the data.

In the early layers, the SDAE focuses on the data's fine-grained features. It learns

to encode fundamental properties like edges, textures, and simple patterns. As you progress deeper into the network, more layers are added on top of it. They abstract and summarize the underlying parts to provide more compact and meaningful representations rather than just copying them. This hierarchical learning is crucial for dimensionality reduction. The initial levels include specific information, whereas the deeper layers capture broader and more abstract concepts. By the time data reaches the SDAE's final layers, it has reduced the core of the input to a compressed form. This reduced-dimensional representation retains the most critical information while reducing unnecessary complications and noise.

In this method, SDAE is a useful tool for unsupervised feature learning and dimensionality reduction. It does not require explicit labels or existing categories; rather, it automatically discovers and organizes the most significant sections of the data. The resulting reduced-dimensional representation is more than just a compressed version of the original input; it is a distilled form that emphasizes the most significant features, making further analysis, visualization, and modeling simpler and faster.

3.4 PCA

Principal Component Analysis (PCA) is a dimensionality reduction technique commonly used in data analysis and machine learning. Its primary goal is to transform a high-dimensional dataset into a lower-dimensional space while preserving as much of the original variability as possible. PCA accomplishes this by identifying the principal components, which are linear combinations of the original features.

Before applying PCA, it's essential to standardize the data to have a mean of 0 and a standard deviation of 1. This ensures that all variables are on the same scale, preventing any particular feature from dominating the analysis due to its larger magnitude. For each observation, subtract the mean and divide by the standard deviation: standardized_data = (data - μ) / σ
The next step is to compute the covariance matrix of the standardized data. Multiply the standardized data matrix by its transpose: covariance _matrix = standardized _data * standardized _data.* T. The covariance matrix provides information about the relationships between different features. It is a symmetric matrix where each element represents the covariance between two variables.

The next Course of Action is to Perform eigendecomposition on the covariance matrix to obtain its eigenvectors and eigenvalues. The eigenvalue problem for the covariance matrix: covariance _matrix * eigenvector = eigenvalue * eigenvector The eigenvectors represent the principal components, and the corresponding eigenvalues indicate the amount of variance captured by each principal component. Sort the eigenvectors in descending order based on their corresponding eigenvalues.

We choose the top k eigenvectors based on the explained variance to retain in the lower-dimensional space. Form a projection matrix by selecting the top k eigenvec-

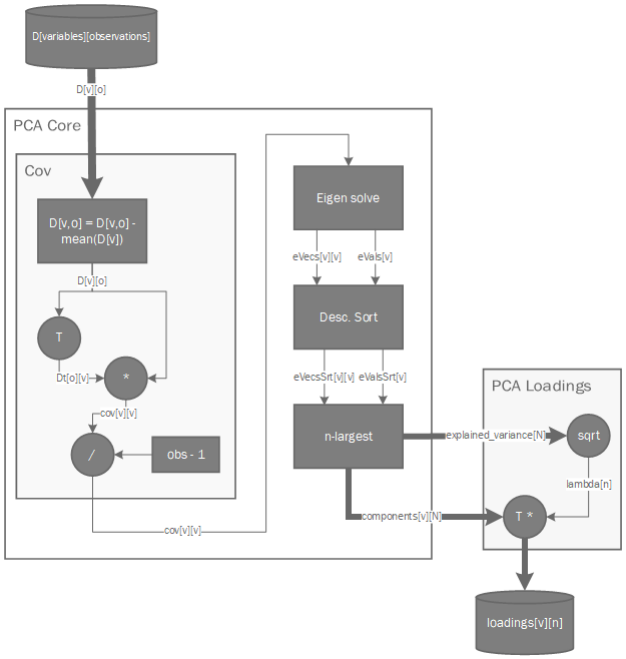

Figure 3.4: Structure Of PCA

tors and arranging them as columns. This matrix is used to transform the original data into the new lower-dimensional space. Multiply the standardized data by the projection matrix to obtain the lower-dimensional representation. The new dataset will have fewer dimensions (features) than the original dataset. The principal components capture the most important patterns and structures in the data. We can examine the loadings of each variable on the principal components to understand their contributions.

3.5 KPCA

Kernel Principal Component Analysis (KPCA) is an extension of Principal Component Analysis (PCA) that allows for nonlinear dimensionality reduction. Unlike PCA, which is based on linear projections, KPCA uses a kernel function to implicitly map the input data into a higher-dimensional space, where linear projections are applied to capture nonlinear relationships.

The choice of the kernel function is crucial in KPCA. Common choices include the radial basis function (RBF) kernel, polynomial kernel, and sigmoid kernel. The kernel function, denoted by K(x,y), computes the dot product in the higher-dimensional space without explicitly mapping the data points.

In KPCA, we need to first compute the kernel matrix K using the chosen kernel function: Kij = K(xi, xj). The Next step is to center the Kernel Matrix K by subtracting the mean of each row, column, and the entire matrix. The centered matrix is denoted by $\vec{K}$

we need to Perform eigendecomposition on the centered kernel matrix K to obtain its

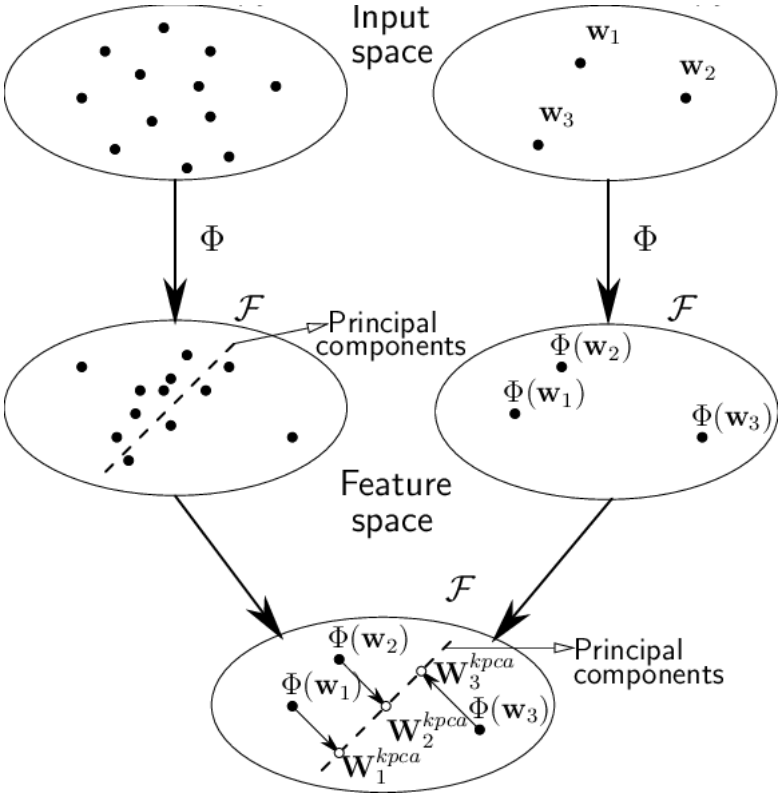

Figure 3.5: Simplified KPCA

eigenvectors and eigenvalues. The eigenvectors represent the principal components in the high-dimensional feature space. We need to Select the top k eigenvectors corresponding to the largest eigenvalues to form the projection matrix (P). We need to project the original data (X) onto the new subspace using the projection matrix (P): Y(Tranpose)*P.The rows of the matrix Y represent the reduced-dimensional representations of the original data in the feature space defined by the selected kernel.

3.6 Classification And Learning

3.6.1 ANN

A key idea in the fields of machine learning and artificial intelligence is artificial neural networks, or ANNs. ANNs are an attempt to mimic how humans learn and absorb information. They are inspired by the neural structure of the human brain. Let's take a narrative look at the idea of ANNs.

Computer scientists always hoped to create devices that might match the brain's amazing ability for experience-based learning. This research produced Artificial Neural Networks, a cutting-edge technology inspired by the complicated network of neurons in the human brain. Considering a huge network of artificial neurons,

or nodes, connected, with each node representing a basic computer unit. Similar to their biological counterparts, these artificial neurons work together to process information. In this network, information travels along channels similar to how electrical impulses travel through the brain's neural circuitry.

The concept of layers is crucial to artificial neural networks. Consider these layers to be different degrees of abstraction that help the network solve problems. The input layer receives information, the hidden levels process it, and the output layer generates the finished product. During the training phase, the network receives a set of tagged data. After the input layer receives the data values, the network uses a learning process to alter its internal parameters. The network's comprehension and ability to distinguish between different types of input increases continuously as information flows via hidden layers. Backpropagation is what makes it all happen. The network acts similarly to how a student receives feedback on their comments. The backpropagation algorithm helps the network detect faults in photo categorization and adjust its parameters accordingly. This approach is continued until the network accurately distinguishes between cats and dogs. Once trained, the Artificial Neural Network can apply its newly acquired knowledge to previously viewed images.

ANN has evolved into a strong tool for a variety of tasks, including picture recognition and natural language processing, as well as game play and event prediction.In the broad tapestry of technology, Artificial Neural Networks are a monument to humanity's desire to reproduce its own cognitive processes. As these networks expand, they have the potential to unlock new levels of intelligence, forever altering the way humans interact with machines and the world around us. And so the story of Artificial Neural Networks begins, a riveting chapter in the unending saga of human ingenuity.

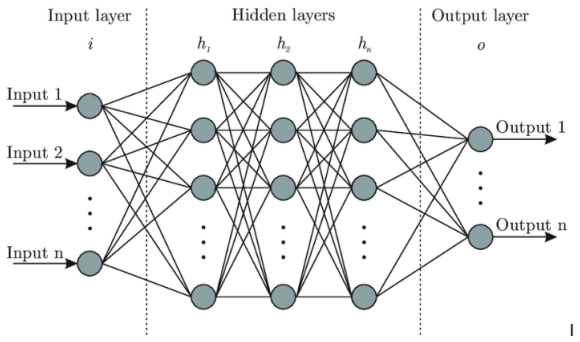

Figure 3.6: Structure Of ANN

An Artificial Neural Network (ANN) was employed in this study to improve our analysis and detect complicated patterns in the high-dimensional gene expression data. ANNs are a type of machine learning model that takes inspiration from the design and operation of the human brain, namely the networked neurons within. The design of our ANN took into account both our study's goals and the characteristics of the genomic data.

Our neural network is made up of various layers, each of which serves a specific purpose during the learning process. The input layer corresponds to gene expression data, and the network's successive hidden layers allow it to learn elaborate representations of the underlying patterns. The final output layer generates a binary classification result that distinguishes between tumor and non-tumor samples. Our custom ANN design, combined with appropriate regularization and hyperparameter tweaking, is a potent tool for identifying complicated correlations in genomic data, resulting in reliable tumor sample categorization.

3.6.2 SVM

Classifying data is a common machine-learning task. The goal is to decide which of the two classes the given data points belong to, and suppose that a new data point needs to be inserted in one of them. In the framework of support vector machines, we wish to know if we can partition data points into (p-1)-dimensional hyperplanes. We call this a linear classifier. Many hyperplanes could be utilized to categorize the data. The hyperplane with the greatest margin, or distance, between the two classes, is a strong candidate for best hyperplane. As a result, we choose the hyperplane in a way that optimizes the distance on each side between it and the nearest data point. A maximum-margin classifier, also known as the perceptron of optimal stability, is a linear classifier defined by a maximum-margin hyperplane if one exists. Support vector machines (SVMs), also known as support vector networks, are supervised max-margin models with learning algorithms that examine data for both classification and regression analysis.

A strong and adaptable machine learning approach for classification and regression applications is called Support Vector Machine (SVM), which was developed by Vladimir Vapnik and his colleagues in the 1990s, has grown in popularity due to its ability to handle both linear and nonlinear correlations between features and target variables. Essentially, SVM seeks the optimum hyperplane in a high-dimensional space to partition data points into discrete classes. This hyperplane functions as a decision boundary in binary classification, categorizing examples on one side as belonging to one class and cases on the other as belonging to the opposite class.

The hyperplane selected by SVM optimizes the margin, which is the distance between the hyperplane and the nearest data point from either class. Maximizing the margin improves the model's generalization performance, making it more robust to new, previously unknown data. The data points that are closest to the hyperplane are known as support vectors, and they are crucial in determining the margin. These support vectors are crucial in locating the optimal hyperplane and contribute to the overall robustness of the model.

When data is not linearly separable, SVM can discover a hyperplane to separate it by mapping the original characteristics into a higher-dimensional space using a kernel technique. The linear kernel is a typical kernel for linear separation, but polynomial and radial basis function (RBF) kernels are frequent for nonlinear separation.

y is the class value of the training samples; $y \in \{1, -1\}$.

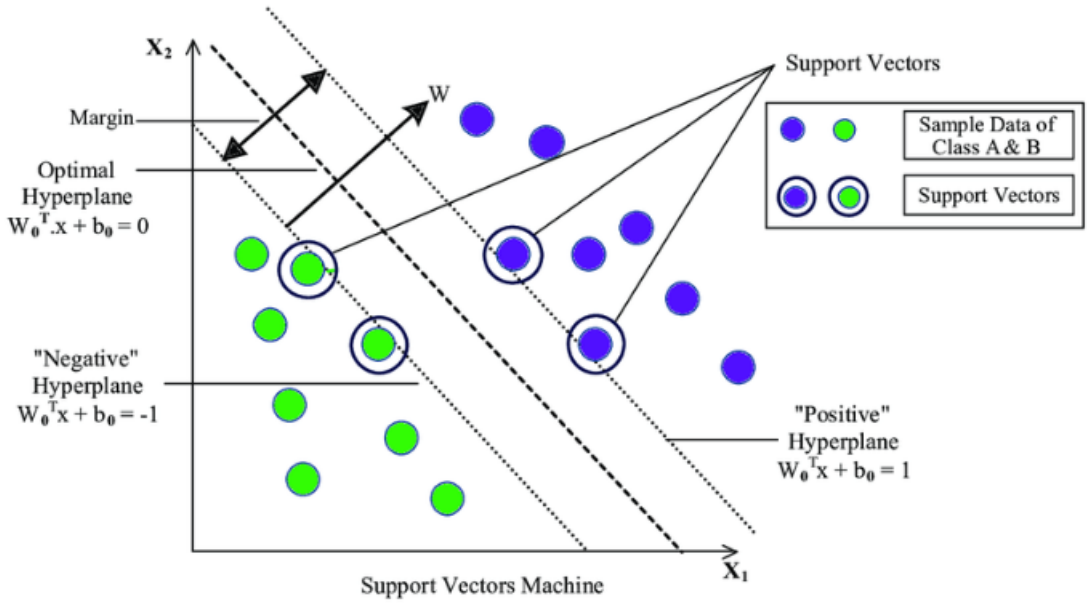

Figure 3.7: SVM Mechanism

SVM penalizes misclassified data points by using a cost function. To get a large margin while avoiding categorization errors, a compromise must be reached. The cost parameter, sometimes known as the letter C, manages the trade-off between achieving a smooth decision boundary and accurately classifying training data.

The SVM offers a lot of advantages. Even when there are more characteristics than samples, SVM performs effectively. SVM is less prone to overfit since it generalizes better to new, previously unseen data by optimizing the margin. Furthermore, the ability to use different kernels enables SVM to handle a wide range of data distributions, both linear and nonlinear. Convex optimization problems must be solved so that SVM can find the global optimum rather than becoming caught in local optima.

Because SVM is effective in high-dimensional domains, it can be applied to problems involving a large number of features. SVM can handle our project's high-dimensional data, which includes a large number of gene expression features, efficiently.

Using the kernel technique, SVM simulates non-linear connections between features and the target variable. The radial basis function (RBF) kernel (kernel='rbf') is particularly useful for modeling complex, nonlinear decision boundaries. This is especially important because linear approaches may not be sufficient to model the relationships in our data.

3.6.3 SVM-RBF

The RBF kernel is a sort of kernel function that maps input features to a higher-dimensional space. The mathematical definition of the RBF kernel is:

$$K(Xi, Xj) = exp(-\frac{|Xi-Xj|^2}{2\alpha^2})$$

Support Vector Machine with Radial Basis Function (RBF) kernel, often referred to simply as SVM RBF, is a variant of the Support Vector Machine algorithm that uses the radial basis function as the kernel function. The RBF kernel allows SVM to handle non-linear relationships between features and target variables, making it suitable for a wide range of complex datasets.

Here, it represents the Euclidean distance between the vectors, and sigma is a parameter that controls the width of the Gaussian distribution.

When the RBF kernel is employed in SVM, the method can implicitly translate the input features to a higher-dimensional space. This change allows the SVM to discover a non-linear decision boundary in the original feature space, making it appropriate for datasets with complicated patterns.

The decision function of an SVM with RBF kernel can be expressed as:

$$f(X) = \sum \alpha iyiK(Xi + X) + b$$

Because SVM can capture complicated relationships in the data thanks to the RBF kernel, it can be used in a variety of situations where the decision boundary is non-linear. The Gaussian distribution's width is set by the RBF kernel's σ parameter. Selecting the right value for σ is important and might need to be adjusted for best results. A smoother decision boundary is produced by high σ values, whereas a more complex border is produced by low σ values. SVM with RBF kernel is a potent tool for classification problems in real-world contexts because it can handle datasets that are not linearly separable.

Chapter 4

Data Analysis

4.1 Dataset Description

Our investigation into breast cancer gene expression encompasses three distinct datasets, each contributing valuable insights into the intricate molecular landscape of this prevalent disease. We get the dataset from a similar work done in [20] where different techniques were employed on these datasets to detect cancer. In this data, the cancer samples were annotated using the patient barcode (uid) for each sample, which was encoded in the variable known as "Hybridization REF." The features contain gene symbols, representing specific genes in the human genome. The sample-specific values for the features are differentially expressed from their normal states done by linear modeling using the R-limma package.

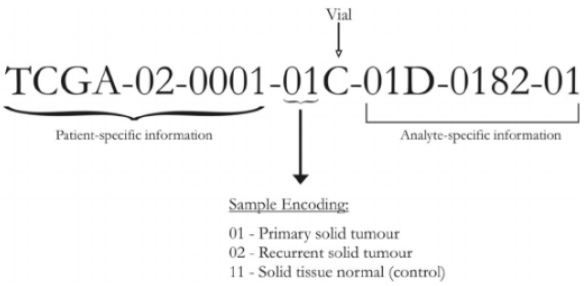

Figure 4.1: TCGA 'Hybridization REF'Barcode

4.1.1 BC-TCGA Dataset:

The BC-TCGA dataset [11] is a thorough compilation that includes 17,814 genes from 590 samples. Within this dataset, we see a clear distinction between normal tissue (61 samples) and breast cancer tissue (529 samples). This comprehensive resource serves as the foundation for our investigation of the many genetic markers connected with breast cancer.

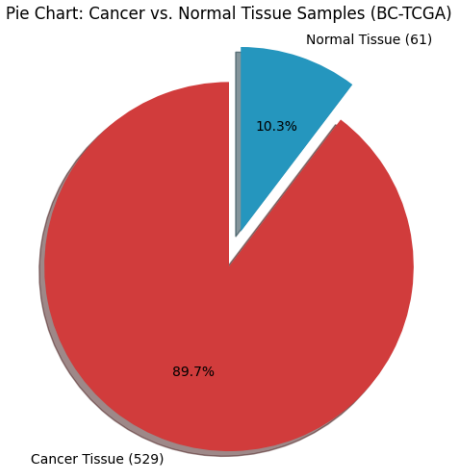

Figure 4.2: Pie chart of sample type (BC-TCGA)

4.1.2 GSE25066 Dataset:

GSE25066 [10], which contains 12,634 genes and 492 breast cancer samples, adds an extra degree of intricacy. This dataset categorizes samples based on their pathological response, with 100 demonstrating a pathologic complete response (PCR) and 392 having residual disease (RD). This dataset's multifaceted character contributes to our understanding of breast cancer heterogeneity and assists in the identification of genes related to therapy response.

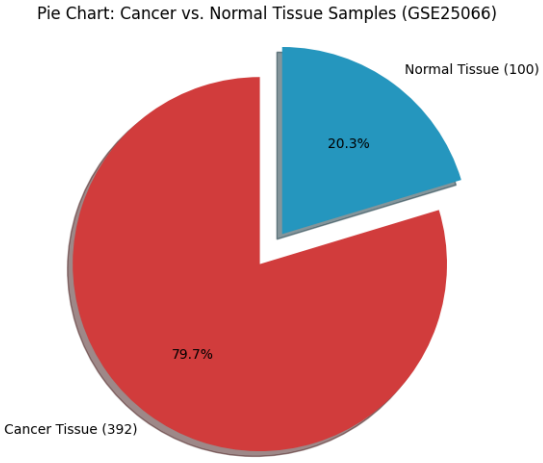

Figure 4.3: Pie chart of sample type (GSE-25066)

4.1.3 Dimensionality Description

Because our data contains many more characteristics or variables than samples, it is classified as high-dimensional. Because of its high complexity, analysis and interpretation may prove difficult. The dimensionality curse exacerbates difficulties such as increased computational complexity and overfitting. To solve these challenges, dimensionality reduction techniques are employed to simplify the data and extract important information, resulting in more accurate and efficient analysis.

A large number of variables in a high-dimensional dataset might cause more noise and redundancy, making it difficult to discern relevant patterns and relationships. Certain dimensionality reduction techniques can help alleviate these issues by translating data into a lower-dimensional space while retaining crucial information. This will simplify computing operations while also improving the interpretability and generalizability of machine learning models employed with the data.

The presence of random or irrelevant information in noisy data makes it difficult to detect patterns and insights. Noise in the data can lead models to perform poorly in machine learning tasks, compromising the algorithms' robustness and ability to generalize. Preprocessing approaches, such as the use of denoising autoencoders or noise reduction techniques, are critical for enhancing data quality and the overall performance of subsequent modeling and analysis.

As we sift through the vast genomic landscapes provided by these datasets, our focus lies on identifying key biomarkers. These molecular indicators could serve as critical tools for early detection and prognosis. The integration of multiple datasets not only reinforces our findings but also allows for a more holistic understanding of the complex molecular underpinnings of breast cancer.

4.2 Data Pre-processing

We started by combining the tumor and normal samples from each dataset. This permitted the development of unified datasets for BC-TCGA, GSE2034, GSE25066, and Simulation Data, allowing for direct comparisons of tumor and normal samples. The datasets were then transposed to make them easier to visualize and analyze and to take them in the correct form to represent features..To obtain impartial results, we shuffled the transposed datasets. Visualizing the gene expressions of 'yes' (tumor) and 'no' (normal) samples revealed different patterns.

To ensure the correctness and integrity of our data, we replaced any missing (NaN) values in the dataset. The gaps created by missing values were filled in with imputed values using a suitable technique, such as mean or median imputation, which kept the data's overall structure. This phase has a significant impact on the robustness of subsequent analyses and the prevention of data loss.

To standardize gene expression values, we then employed Z-score normalization. A z-score indicates the number of standard deviations above or below the mean of a data point. Here's the formula to calculate a z-score: z = data point − mean standard deviation($z = x - \mu * \sigma$). By reducing scale disparities between genes, this normalizing process ensures that each gene contributes equally to the analysis. Z-score normalization eliminates bias caused by varying scales among genes by transforming the data to a single scale with a mean of 0 and a standard deviation of 1. This facilitates fair comparisons and the identification of significant trends. We did these tasks for each of the separate datasets.

This concludes our data preprocessing and now with its results, we continue to

apply other necessary methods to extract features, reduce dimensionality and deal with noise in data. Our narrative encompasses a thorough examination of various datasets, each of which adds a unique perspective to our overall understanding of breast cancer. The combination of empirical and simulated data enhances our analytical method, paving the path for significant contributions to breast cancer research and tailored medicine.

Chapter 5

Methodology

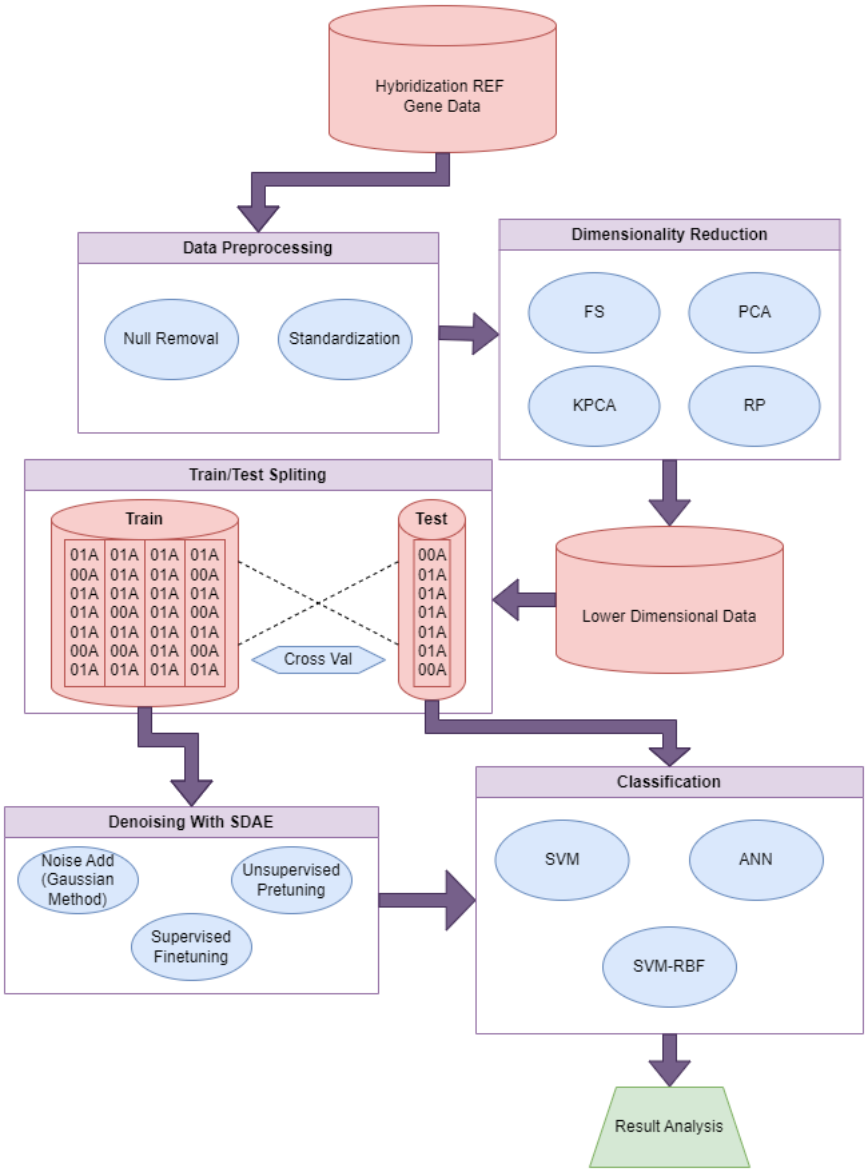

Figure 5.1: Workplan of the methodology

5.1 Dimensionality Reduction and Denoising

In the first step of our method, we minimize the complexity of the story dataset using Dimensionality Reduction approaches. This stage is critical for extracting important features with minimal information loss. For dimensionality reduction, we have employed

5.1.1 Feature Selection

In our project, we employed univariate feature selection. One of the filtered feature selection methods uses the chi-squared test. Methods for selecting univariate features evaluate each feature independently about the target variable, ignoring feature interactions. These strategies select the highest-ranked characteristics using statistical tests or other criteria. The chi-square (X^2) test assesses the relationship between two categorical variables by comparing their predicted and observed co-occurrence frequencies. The chi-square test is used in the feature selection process to assess the relationship between each characteristic and the target variable, assisting in the identification of the most important features for prediction.
Equation for Chi-Square Test:

$$X^2 = \sum \frac{(Oi-Ei)^2}{Ei}$$

Where:
Oi = Represents the frequency that is observed in the dataset for a specific category.
Ei = The frequency that is expected if there were no relationships between the variables, calculated assuming independence between the variables.

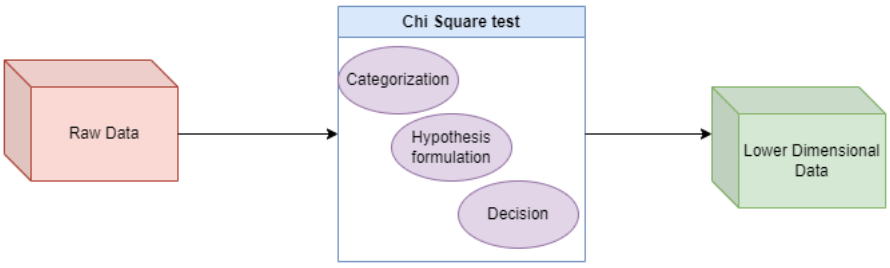

Figure 5.2: Feature Selection(Chi-square)

5.1.2 Principal Component Analysis

Principal Component Analysis (PCA) is an important tool in our work on dimensionality reduction. This method attempts to preserve significant variance in the data while transforming the original feature space to a lower-dimensional representation. Using principal component analysis (PCA), we can decrease the amount of

data without losing essential patterns by finding which principal components account for the majority of the dataset's variability. In this phase, we apply PCA to accurately capture the fundamental structure of our dataset. We are using the reduced principle component set as the foundation for future study, allowing us to examine our dataset more efficiently and specifically.

Procedures in Principal Component Analysis:

1. Normalize the data: Since PCA is affected by the magnitude of the variables, the first step is to normalize the dataset to guarantee that the average of each feature is 0 and the standard deviation is 1.

2. Covariance matrix: We calculate the covariance matrix for the analysis of variable relationships. The covariance matrix assists in identifying correlation patterns within the dataset.

3. Calculate the eigenvectors of the principal components and the eigenvalues: the covariance matrix and the direction of the principal components are defined by the eigenvectors, whereas the eigenvalues represent the extent of variance accounted for by each principal component.

4. Arranging and selecting primary components: organize the eigenvectors according to their eigenvalues in a descending order. Select the primary eigenvectors that denote the greatest variability within the dataset. These are the primary elements employed for dimensionality reduction.

5. Modify the data: We project the original data onto the newly established space defined by the selected principal components, thereby reducing the dimensionality.

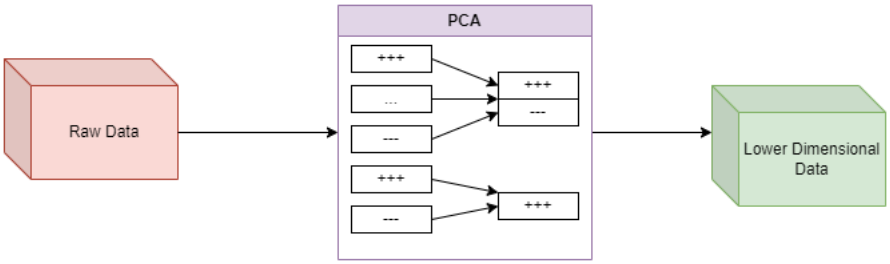

Figure 5.3: Principle component Analysis

PCA Formula:
PCA's mathematical basis is centered on eigenvalues and eigenvectors. Data can be converted into principal components through a process known as transformation.

$$Z = XW$$

Where,
Z = The Transposed data matrix
X = The original data matrix
W = The eigenvector matrix

The principal components are obtained by solving the eigenvalue problem for the covariance matrix C

$$CW = \lambda W$$

Where,
C = The covariance matrix
λ= The diagonal matrix of eigenvalues
W = The eigenvector matrix

5.1.3 KPCA

We employed Kernel Principal Component Analysis (KPCA) to reduce the dimensionality of the high-dimensional gene expression data. KPCA maps the data into a more separable space by capturing complex correlations between genes with nonlinear kernels such as RBF. By focusing on the most important properties for cancer detection, KPCA improves the efficacy and accuracy of subsequent classifiers like SVM and ANN by reducing the amount of features while preserving crucial nonlinear patterns.

Procedures in Kernel Principal Component Analysis:

1. Mapping data into Higher dimensional space: The KPCA uses the kernel function to transform the original data into higher dimensional space, where it is easier to find the principal component. Instead of directly computing the change to a higher-dimensional space (which can be costly), KPCA utilizes a kernel function to determine the inner product of the data points in the same high-dimensional space. This kernel trick eliminates the requirement for physical transformation.

2. Calculating kernel Matrix:: A kernel matrix K is computed, where each element $Kij = k(xi, xj)$ represents the kernel value (similarity) between two data points xi and xj.

3. Problem of eigenvalues in kernel space: instead of utilizing the covariance matrix, KPCA identifies the eigenvalues and eigenvectors within the kernel matrix. The eigenvectors represent the directions of maximum variance in a multidimensional space, whereas the eigenvalues quantify the amount of variance represented by each principal component.

4. The data is transformed by projecting the original data points onto principal components within the kernel-transformed space. This results in a reduction of dimensionality by integrating nonlinear elements that encompass complex relationships within the data.

KPCA Equation: The transformation in KPCA is represented as:

$$K(Xi, Xj) = \phi(Xi)^T \phi(Xj)$$

Where,
$K(Xi, Xj)$ = The data point Xi and Xj
ϕ = Mapping into higher dimensional space.

5.1.4 SDAE

The Stacked Denoising Autoencoder (SDAE) enhanced with Gaussian noise is a critical component of our denoising process. SDAE is used as a powerful unsupervised learning method, capable of collecting complex hierarchical structures inside data. By incorporating Gaussian noise into the training process, the SDAE learns to reconstruct the clean, underlying characteristics while filtering out unwanted noise. This denoising mechanism increases the model's robustness and makes future representations more resilient to input variations. Using SDAE with Gaussian noise, our approach aims to reveal hidden patterns in our non-narrative dataset, resulting in a refined and noise-resistant feature space for subsequent analysis.

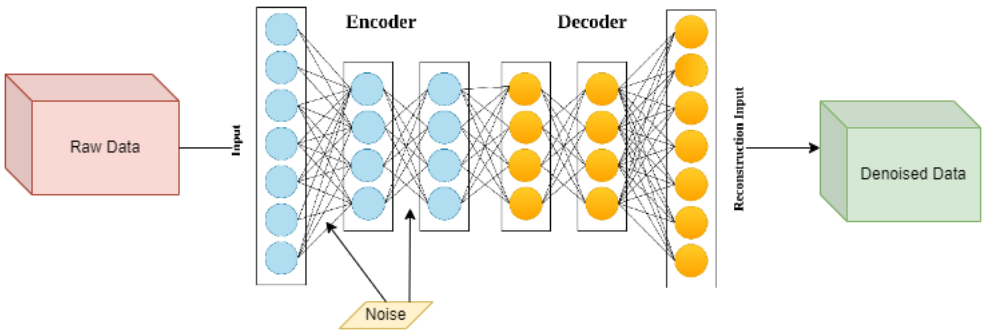

Figure 5.4: The Stacked Denoising Autoencoders

SDAE Equation: The primary goal of the autoencoder is to reduce the reconstruction error:

$$L(x, \hat{x}) = ||x - \hat{x}||^2$$

Where,
x = The original data
$\hat{x}$ = The reconstructed data from the autoencoder

For SDAE, the input is corrupted by the addition of Gaussian noise:

$$\hat{x} = x + N(0, \sigma^2)$$

Where,
x = The original data
$\hat{x}$ = The noisy input
$N(0, \sigma^2)$ = Gaussian mean 0 and variance σ

5.1.5 FS+SDAE

We describe a unique approach for our dataset that combines Stacked Denoising Autoencoders (SDAE) with Feature Selection (FS). To reduce processing complexity and potential noise, feature selection is critical for choosing the most informative variables for subsequent analysis. We will employ techniques such as recursive feature elimination and mutual information with the help of the chi-square test which chooses relevant features corresponding to our target variable tumor'. The chosen features are then utilized to train a Stacked Denoising Autoencoder architecture, which will achieve two objectives: dimensionality reduction and noise removal. With the chosen gestures from FS, SDAE then starts adding noise to the data in each layer for training. By refining the dataset's representation, this hybrid FS+SDAE framework aims to increase the performance of our analytical models by getting meaningful features as well as being able to deal with noisy and corrupt data. This therefore ensures greater efficiency in the classification of our target.

Equation (Combined FS + SDAE Process):

Feature Selection: Let Xselected be the feature set selected by FS from the original dataset X.

$$\text{Xselected} = FS(X)$$

SDAE Reconstruction: After adding noise $N(0, \phi^2)$ to Xselected the noisy data Xselected is fed into the SDAE, which reconstructs the clean features:

$$\hat{X}selected = Xselected + N(0, \sigma^2)$$

The autoencoder then minimizes the reconstruction loss:

$$L(Xselected, \hat{X}Selected) = ||Xselected - \hat{X}selected||^2$$

5.2 Evaluation Models:

5.2.1 SVM

The Support Vector Machine (SVM) is an important component of our strategy, as it aids in dimensionality reduction and classification. The most discriminative features are discovered and extracted using SVM from the enhanced feature space obtained from the dimensionality reduction and denoising steps. SVM helps to provide a compact representation of data that highlights class boundaries by maximizing the margin between classes. Furthermore, SVM serves a second function as an assessment model, providing a credible indicator of classification performance. Our goal is to increase classification accuracy on our non-narrative dataset by integrating SVM, denoising, and dimensionality reduction techniques. This allows for a detailed and efficient study of the underlying patterns.

5.2.2 ANN

Our neural network is made up of various layers, each of which serves a specific purpose during the learning process. The input layer corresponds to gene expression data, and the network's successive hidden layers allow it to learn elaborate representations of the underlying patterns. The final output layer generates a binary classification result that distinguishes between tumor and non-tumor samples. Our custom ANN design, combined with appropriate regularization and hyperparameter tweaking, is a potent tool for identifying complicated correlations in genomic data, resulting in reliable tumor sample categorization.

5.2.3 SVM-RBF

The Support Vector Machine with Radial Basis Function (SVM-RBF) plays an important part in our proposed methodology, providing a sophisticated approach to nonlinear classification. SVM-RBF uses the kernel method to effectively translate input data into a higher-dimensional space, making it easier to identify detailed patterns in our non-narrative dataset. In terms of dimensionality reduction, SVM-RBF excels at capturing complicated relationships between features, resulting in a better representation of the underlying data structure. Furthermore, SVM-RBF functions as an integral evaluation model, helping to provide a full assessment of classification performance. The careful incorporation of SVM-RBF into our technique aims to reveal nuanced insights and strengthen the research on our unique dataset.

5.3 Experimental Setup

For our experimental setup, we focused on three gene expression datasets and by following a rigorous preparation step, dimensionality reduction, and classification, each dataset was reviewed independently.

The preprocessing procedure began with scaling the data using MinMaxScaler or StandardScaler to provide consistent feature ranges across the gene expressions. A range of dimensionality reduction techniques, which are essential for improving model accuracy and computational efficiency, were subsequently used to convert the data into more manageable feature spaces. By employing dimensionality reduction strategies like Principal Component Analysis (PCA), Random Projection (RP), and feature selection strategies like SelectKBest (using chi-squared scoring), we were able to condense the initial high-dimensional space (17,816 genes) into smaller, more pertinent feature subsets. We were able to remove noise from the data and maintain the most significant patterns by employing these strategies. For each reduced dataset,

three classification models were trained: an Artificial Neural Network (ANN), a Support Vector Machine (SVM) with a linear kernel, and an SVM-RBF kernel. When combined with a complex feature extraction process, such as stacked denoising autoencoders (SDAE), the SVM-RBF kernel's capacity to detect non-linear patterns in the data improves dramatically. SDAE was added to the pipeline to capture nonlinear structures and improve feature representation by denoising and reconstructing the input features. To ensure accurate performance assessment, each model under-

went tenfold stratified cross-validation. For gene expression data with often uneven tumor vs. non-tumor labels, stratification helped to maintain the balance of classes in each fold. Following training on the shortened feature sets, we evaluated each

model's performance using the standard metrics of accuracy, precision, recall, and F1 score. These measurements provided insight into the model's overall classification accuracy as well as its ability to control class imbalances.

Chapter 6

Result Analysis

The culminating findings of our extensive testing shed insight into the performance of various dimensionality reduction and denoising approaches when combined with different assessment models. Notably, the combination of Feature Selection (FS) and Stacked Denoising Autoencoders (SDAE) emerges as a noteworthy strategy, achieving outstanding accuracy across several assessment models, on all our datasets.

For the first dataset which holds the regular samples, the FS+SDAE approach combined with Support Vector Machine Radial Basis Function (SVM-RBF) achieves a remarkable average accuracy of 99.15%, with precision, recall, and F1 scores over 99%. It also achieves high scores in all evaluation metrics with the other models and has over 99% score in recall for all the classification models. This shows how feature selection and denoising work together to improve the robustness of classification problems.

Furthermore, when paired with SVM, Principal Component Analysis (PCA) effectively reduces dimensionality, resulting in high accuracy ratings. However, when paired with SVM-RBF, accuracy significantly declines, demonstrating sensitivity to assessment model selection. Stacked Denoising Autoencoders (SDAE) with Gaussian noise and Random Projections (RP) paired with SDAE produce comparable results, highlighting the adaptability of denoising approaches. While the SVM-RBF performs differently depending on the technique, it has consistently demonstrated the ability to capture complicated patterns.

When both the 2nd and 3rd dataset's results are analyzed, we find slight improvements in accuracy though it is not much. However, when the recall value we have managed to get is extremely high which is the most important metric in the case of cancer detection as it minimizes the chances of detecting false negatives. In many of the cases, mostly when the SVM-RBF model is run, the recall score is over 90%.

The reason for SVM-RBF to have such a good performance, in this case, would be because of underfitting in the main model. SVM using an RBF kernel solves underfitting by introducing a nonlinear decision boundary. The RBF kernel converts the input data into a higher-dimensional space, allowing the SVM to distinguish data points that were not linearly separable in the original feature space. This increases the model's capacity to detect more complex patterns in the data, reducing

underfitting when simpler linear models are unable to understand the underlying structure. By adjusting the hyperparameters, the SVM-RBF may successfully balance complexity and generalization.

Notably, we find excellent results for the SVM-RBF, when denoising is performed with SDAE before running it. It is mostly because SDAE organizes the latent space by denoising the data and identifying strong, non-linear properties. This update highlights complex data links that would otherwise be difficult to discern in raw characteristics. Using this new feature space, SVM-RBF, which is designed to capture nonlinear patterns, successfully differentiates classes in a high-dimensional environment. SVM-RBF can increase classification performance by employing SDAE's denoised and refined features, especially when dealing with complicated patterns like gene expression data.

It is vital to remember that the methods for dimensionality reduction and denoising have a significant impact on the outcomes, thus selecting carefully is critical to get the best results. The FS+SDAE methodology is strong and adaptable, and its results are promising enough to warrant further exploration in non-narrative datasets. These findings provide valuable information for academics and practitioners seeking efficient techniques for preparing and assessing data for categorization jobs. Moving on, we want to expand our research with more combinations of advanced techniques to see if the efficiency of detecting cancer by gene expressions can be further improved.

6.1 Result of dataset BC-TCGA

Feature	Model	Accuracy	Precision	Recall	F1-Score
FS	SVM	98.12	0.996	0.996	0.996
	ANN	98.19	0.989	0.994	0.997
	SVM-RBF	98.32	0.986	0.996	0.996
PCA	SVM	98.83	0.948	0.998	0.979
	ANN	**98.98**	0.996	0.988	0.994
	SVM-RBF	89.66	0.896	0.9831	0.945
KPCA	SVM	97.08	0.97	0.997	0.983
	ANN	95.48	0.952	0.925	0.975
	SVM-RBF	89.6	0.896	0.983	0.945

Table 6.1: Comparison of Dimensionality reduction using three classification models (BC-TCGA Dataset)

Feature	Model	Accuracy	Precision	Recall	F1-Score
FS+SDAE	SVM	99.32	0.9981	0.9943	0.9961
	ANN	**99.49**	0.998	**0.994**	0.997
	SVM-RBF	99.19	0.998	0.996	0.998
PCA+SDAE	SVM	99.15	0.996	0.994	0.995
	ANN	99.49	0.998	0.99	0.994
	SVM-RBF	89.66	0.896	1	0.945
KPCA+SDAE	SVM	98.32	0.996	0.996	0.996
	ANN	99.49	0.996	0.992	0.994
	SVM-RBF	89.66	0.896	0.998	0.945

Table 6.2: Comparison of dimensionality reduction and denoising using three classification models (BC-TCGA Dataset)

6.2 Result of Dataset GSE-25066

Feature	Model	Accuracy	Precision	Recall	F1-Score
FS	SVM	78.85	0.839	0.908	0.8725
	ANN	80.89	0.853	0.918	0.884
	SVM-RBF	80.48	0.815	0.976	0.888
PCA	SVM	79.87	0.832	0.936	0.881
	ANN	74.99	0.834	0.856	0.844
	SVM-RBF	79.67	0.796	0.998	0.886
KPCA	SVM	79.47	0.796	0.997	0.885
	ANN	80.69	0.810	0.989	0.890
	SVM-RBF	79.67	0.796	0.998	0.888

Table 6.3: Comparison of Dimensionality using three classification models (GSE-25066 Dataset)

Feature	Model	Accuracy	Precision	Recall	F1-Score
FS+SDAE	SVM	74.18	0.837	0.839	0.838
	ANN	80.08	0.848	0.913	0.879
	SVM-RBF	79.27	0.799	0.987	0.883
PCA+SDAE	SVM	73.77	0.838	0.831	0.834
	ANN	77.85	0.848	0.880	0.863
	SVM-RBF	79.67	0.796	0.98	0.886
KPCA+SDAE	SVM	79.47	0.797	0.994	0.885
	ANN	77.84	0.836	0.897	0.865
	SVM-RBF	79.67	0.796	0.998	0.886

Table 6.4: Comparison of dimensionality reduction and denoising using three classification models (GSE-25066 Dataset)

6.3 Confusion Matrices

6.3.1 Confusion matrix of Dataset BC-TCGA:

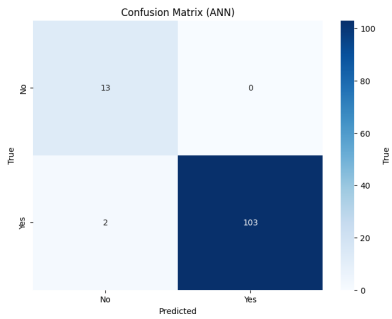

Figure 6.1: FS+SDAE-ANN

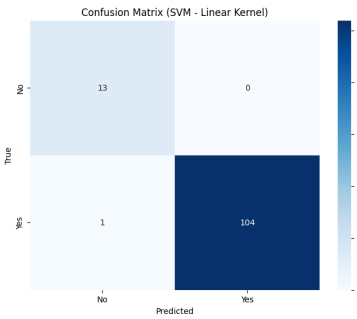

Figure 6.2: FS+SDAE SVM

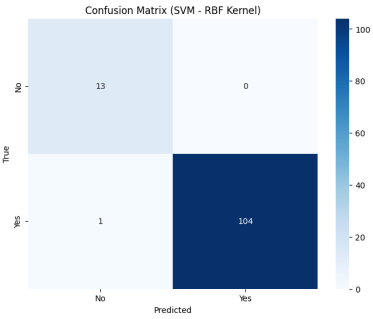

Figure 6.3: FS+SDAE SVM-RBF

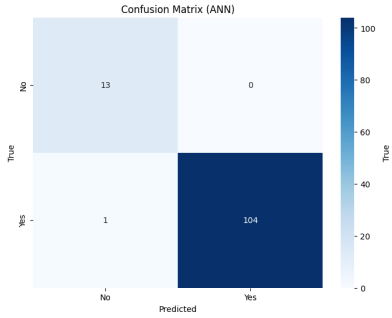

Figure 6.4: PCA+SDAE -ANN

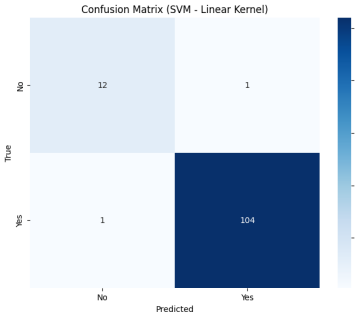

Figure 6.5: PCA+SDAE SVM

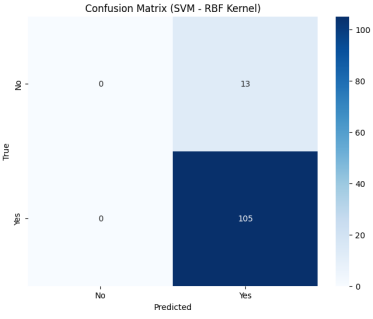

Figure 6.6: PCA +SDAE SVM-RBF

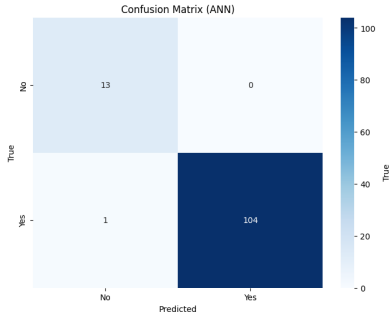

Figure 6.7: KPCA+SDAE -ANN

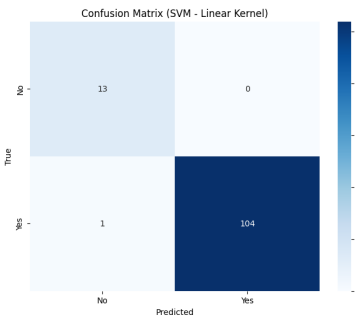

Figure 6.8: KPCA+SDAE SVM

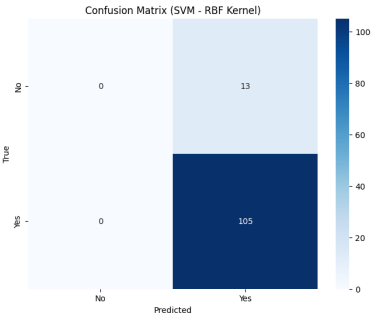

Figure 6.9: KPCA +SDAE SVM-RBF

6.3.2 Confusion matrix of Dataset GSE-25066:

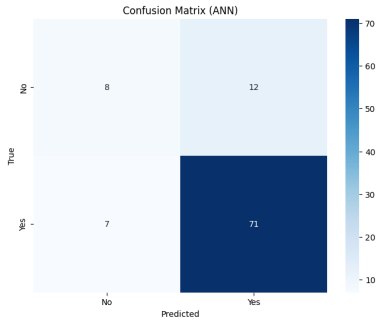

Figure 6.10: FS+SDAE-ANN

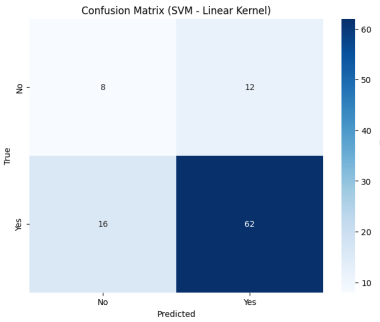

Figure 6.11: FS+SDAE SVM

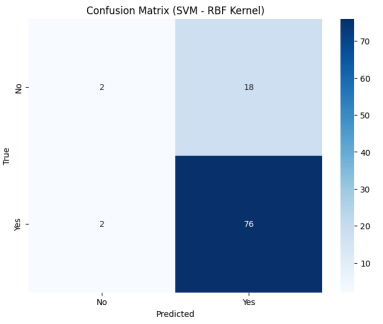

Figure 6.12: FS+SDAE SVM-RBF

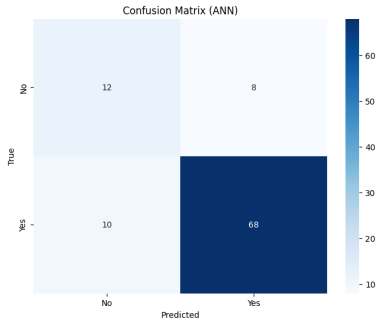

Figure 6.13: PCA+SDAE -ANN

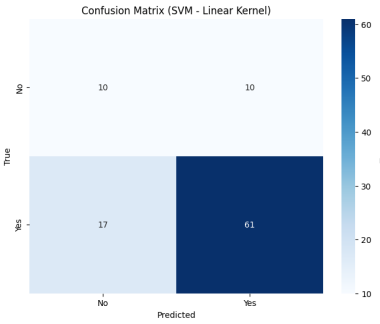

Figure 6.14: PCA+SDAE SVM

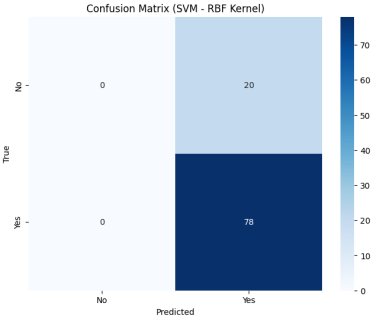

Figure 6.15: PCA +SDAE SVM-RBF

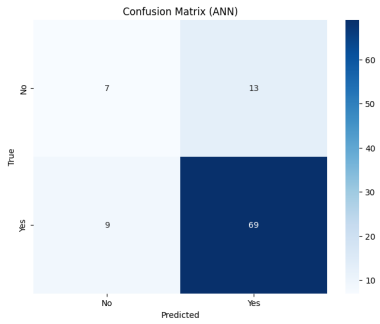

Figure 6.16: KPCA+SDAE - ANN

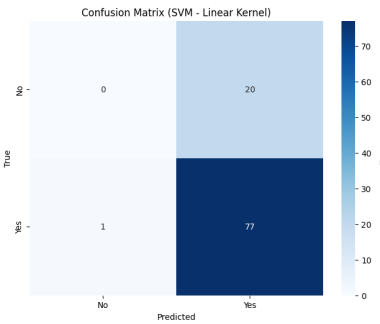

Figure 6.17: KPCA+SDAE SVM

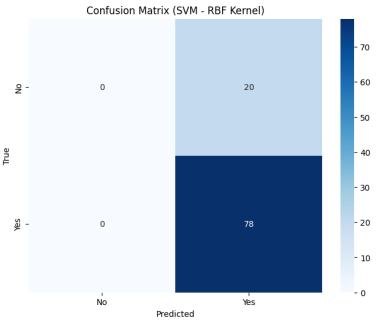

Figure 6.18: KPCA +SDAE SVM-RBF

6.3.3 Confusion Matrix Analysis:

In our study, we leveraged confusion matrices to provide a detailed breakdown of the classification performance for each model. The matrix allows us to evaluate not only the overall accuracy but also the nature of the errors the model makes. Each confusion matrix highlights the relationship between true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This detailed perspective is crucial in our context, as minimizing false negatives is paramount in cancer detection.

For example, using the SVM-RBF model, the confusion matrix in one of our datasets (such as GSE-2034) has a recall score of more than 90%. With extremely few false negatives, this means that the majority of genuine cancer cases were successfully discovered, which is a critical achievement in ensuring timely detection. We modified the model to balance accuracy and recall in this region, however the matrix still demonstrates a trade-off with false positives. We enhanced the models for real-world applications through these tests, ensuring consistent performance across datasets.

6.4 Loss Curve:

6.4.1 Loss curve for BC-TCGA (Dataset 1):

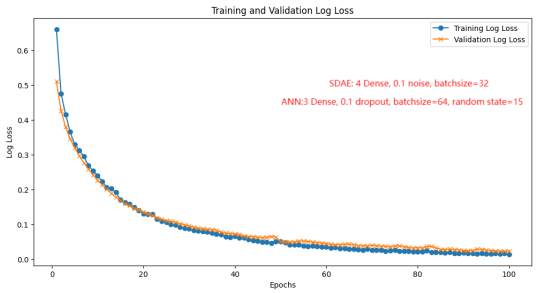

Figure 6.19: Loss-curve of FS + SDAE of BC-TCGA dataset

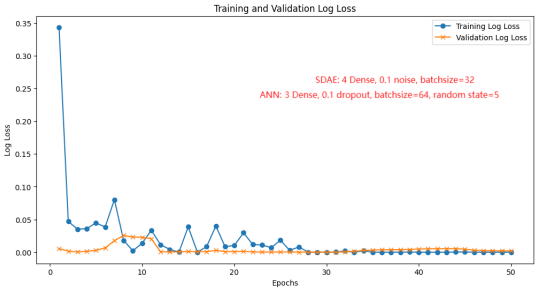

Figure 6.20: Loss-curve of PCA + SDAE of BC-TCGA dataset

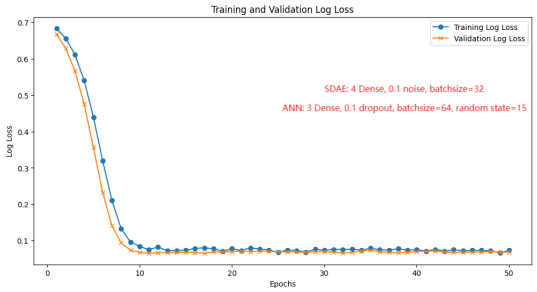

Figure 6.21: Loss-curve of KPCA + SDAE of BC-TCGA dataset

6.4.2 Loss curve for GSE-25066 (Dataset 2):

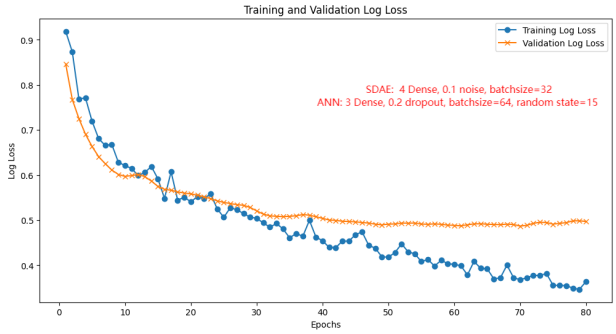

Figure 6.22: Loss-curve of FS + SDAE of GSE-25066 dataset

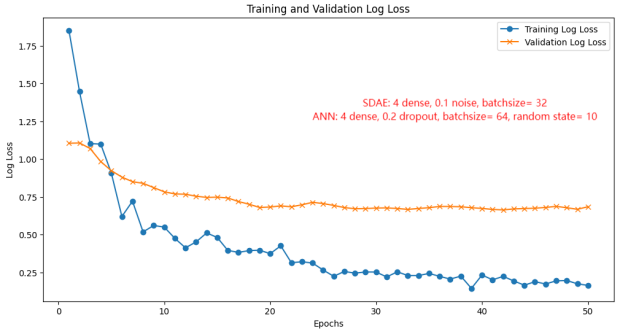

Figure 6.23: Loss-curve of PCA + SDAE of GSE-25066 dataset

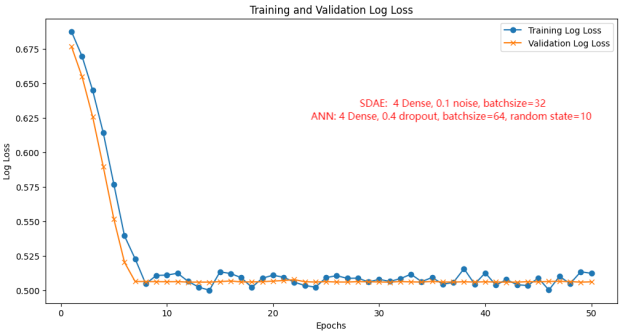

Figure 6.24: Loss-curve of KPCA + SDAE of GSE-25066 dataset

6.4.3 Loss Curve Analysis:

The loss curves given for the first dataset illustrate our models' learning dynamics throughout training and validation. These curves illustrate the training and validation loss over numerous epochs, providing important insights into the model's convergence behavior. These numbers show that training and validation losses have consistently decreased, indicating effective learning.

We observed that, in certain circumstances, especially when employing the SVM-RBF model in combination with SDAE, validation loss plateaued or slightly rose at the end, although training loss decreased. This leads to a condition known as mild overfitting, in which the model becomes overly fitted to the training set. To overcome this, we increased generalization by halting early and modifying hyperparameters. As a consequence, the loss curves assisted us in refining our models and ensuring that they performed well on previously untested training and test datasets.

Chapter 7

Epilogue:

7.1 Limitation:

The project's limitations reflect the project's complexity and the inherent difficulty of utilizing gene expression data for cancer detection. The results demonstrate some limitations that must be acknowledged, even though the results from the various models—particularly the combination of Feature Selection (FS) and Stacked Denoising Autoencoder (SDAE)—were encouraging and mostly achieved impressive performance metrics such as 99.15% accuracy in successful detection for first-time testing patients. However, the GSE25066 datasets suffer from a considerable imbalance in performance metrics, particularly low accuracy despite high recall scores. The models consistently produced high recall scores but lower accuracy in the GSE25066 dataset, which classifies samples based on their pathological response (complete response versus residual disease). This means that the models struggle with overall classification accuracy even while they are effective at finding actual positive cases, which is critical in a medical scenario where failing to detect a cancer diagnosis is quite damaging. According to the high recall but poor accuracy, the models may be misclassifying a significant proportion of non-cancerous samples or samples with no recurrence or residual sickness, increasing the chance of false positives while limiting false negatives. In therapeutic contexts, this mismatch may lead to unnecessary treatments or therapies for patients with incorrect diagnoses, which can be detrimental. Further model tuning or investigation of other approaches to improve overall accuracy while maintaining key recall performance would be necessary to resolve this problem.

Furthermore, the project's biological interpretation is restricted. Although machine learning models, particularly deep learning approaches such as ANN and SDAE, can achieve high accuracy, it is difficult to identify the genes or interactions that generate the predictions due to their "black-box" nature. This lack of interpretability is a major issue in biomedical research, as understanding the biological relevance of observed patterns is just as important as achieving high classification accuracy.

Another limitation is the limited number of samples compared to high-dimensional data. Although the models perform well, particularly the FS+SDAE combination with SVM-RBF, the small sample size (590 samples) may result in overfitting, especially in deep learning models like SDAE and ANN. SVM-RBF's outstanding

performance, in particular, might be attributable to underfitting in the main model since it introduces a nonlinear decision boundary that corrects for underfitting by translating the data into a higher-dimensional space. This demonstrates that, while the findings are positive, the model's generalizability may be limited in the absence of larger datasets or more model fine-tuning.

Another limitation is the dataset's high dimensionality, which comprises 17,816 features (genes) among 590 samples. While approaches such as FS and SDAE are used to reduce dimensionality and noise, they run the risk of removing physiologically important data. The challenge with such high-dimensional data is that certain important genes may be missed in the limited feature space, lowering cancer detection accuracy. This is illustrated by the fact that dimensionality reduction methods like PCA and KPCA provide different outcomes across models. For example, whereas PCA combined with SVM produces excellent results, it performs poorly with SVM-RBF, suggesting sensitivity to the dimensionality reduction strategy and classification model coupling.

7.2 Future Work:

Our approach to using the proposed methods proved to be very effective in the case of applying multiple different deep learning models on two of the datasets yet, there exists some limitations to our work. Thus, our future work shall consist of overcoming such limitations and enriching the flexibility and performance of our methodology beyond what we have achieved so far.

Firstly, one of the most important aspects of the entire research that shall be covered is relevant gene identification for the methodology we have worked on. So far we have used the samples through our techniques but choosing specifically the genes that are more important than others in detecting cancer cells. Selecting and using only the most relevant genes would significantly reduce the work pressure and complexity of the models. We need to choose the most efficient techniques and models to detect the most relevant gene sequences for breast cancer detection.

Secondly, in our second dataset GSE-25066, we couldn't obtain the required results we were hoping for. The accuracy and F1 scores are much less compared to the results from the first dataset BC-TCGA. We have to overcome this issue by reviewing our dimensionality techniques, and efficient model building, making our methodology more versatile over various datasets and still able to gain improved results. Because the current dimensionality reduction techniques have a high risk of removing important gene data that are essential for the overall accuracy of the system.

7.3 Concluison:

To sum up, we worked on creating a strong deep-learning model for cancer diagnosis that makes use of stacked denoising autoencoders (SDAE). We conducted a systematic evaluation of the performance of different dimensionality reduction techniques, namely: Feature Selection (FS), Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPCA), and SDAE, using two different sets of breast cancer gene expression datasets. Following that, these methods were combined with artificial neural networks (ANN), support vector machines (SVM), and SVM with radial basis function (RBF) kernel for machine learning models.

The results of our thorough investigation showed that each applied approach produced different accuracy and classification metric outcomes. Interestingly, the FS + SDAE combination showed remarkable accuracy on all three classification models (SVM, SVM-RBF, and ANN). This emphasizes the usefulness of merging feature selection and denoising autoencoders in increasing the performance of cancer detection models.

Early detection is essential for effective outcomes in the complicated and resource-intensive procedures of cancer diagnosis and treatment. Finding cancer biomarkers from gene expressions and medical imaging has been considerably simplified by the combination of machine learning and deep learning techniques. Notwithstanding these developments, problems like the scarcity of available datasets and the requirement for efficient denoising still exist.

The application of SDAE in combination with dimensionality reduction methods proved to be a very effective strategy for denoising raw data, allowing the extraction of specific features from genome sequence data. We wanted to make cancer detection more effective by avoiding a moderate amount of noise through the encoder model.

More broadly, we believe our work has a major impact on the field of cancer research and diagnosis.As we continue to refine and expand our methods, we hope that our results will provide valuable insights that will ultimately drive advances in early cancer detection and improve patient outcomes.

Bibliography

[1] C. Aliferis, I. Tsamardinos, P. Massion, A. Statnikov, N. Fananapazir, and D. Hardin, "Machine learning models for classification of lung cancer and selection of genomic markers using array gene expression data," Jan. 2003, pp. 67–71.

[2] Y. Wang, J. Klijn, Y. Zhang, *et al.*, "Wang y, klijn jg, zhang y, sieuwerts am, look mp, yang f, talantov d, timmermans m, meijer-van gelder me, yu j, jatkoe t, berns em, atkins d, foekens jagene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. lancet 365: 671-679," *The Lancet*, vol. 365, pp. 671–679, Feb. 2005. DOI: 10.1016/S0140-6736(05)70933-8.

[3] O.-P. Alho, H. Teppo, P. Mäntyselkä, and S. Kantola, "Head and neck cancer in primary care: Presenting symptoms and the effect of delayed diagnosis of cancer cases," en, *CMAJ*, vol. 174, no. 6, pp. 779–784, Mar. 2006.

[4] K. Doi, "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4, pp. 198–211, 2007, Computer-aided Diagnosis (CAD) and Image-guided Decision Support, ISSN: 0895-6111. DOI: https://doi.org/10.1016/j.compmedimag.2007.02.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895611107000262.

[5] L. Fass, "Imaging and cancer: A review," *Molecular Oncology*, vol. 2, no. 2, pp. 115–152, 2008, ISSN: 1574-7891. DOI: https://doi.org/10.1016/j.molonc.2008.04.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574789108000598.

[6] Y. Wang, D. Miller, and R. Clarke, "Approaches to working in high-dimensional data spaces: Gene expression microarrays," *British journal of cancer*, vol. 98, pp. 1023–8, Apr. 2008. DOI: 10.1038/sj.bjc.6604207.

[7] H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," Jan. 2009, pp. 1096–1104.

[8] H. Jefferies, "The lived experience of younger women with vulval cancer: The impact of delayed diagnosis.," *Nursing times*, vol. 106 6, pp. 21–4, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:23729743.

[9] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010, ISSN: 1532-4435.

[10] C. Hatzis, L. Pusztai, V. Valero, *et al.*, "A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer," en, *JAMA*, vol. 305, no. 18, pp. 1873–1881, May 2011.

[11] Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," en, *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012.

[12] M. L. Tørring, M. Frydenberg, W. Hamilton, R. P. Hansen, M. D. Lautrup, and P. Vedsted, "Diagnostic interval and mortality in colorectal cancer: U-shaped association demonstrated for three different datasets," *Journal of Clinical Epidemiology*, vol. 65, no. 6, pp. 669–678, 2012, ISSN: 0895-4356. DOI: https://doi.org/10.1016/j.jclinepi.2011.12.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089543561100388X.

[13] J. Xu, L. Xiang, R. Hang, and J. Wu, "Stacked sparse autoencoder (ssae) based framework for nuclei patch classification on breast cancer histopathology," Apr. 2014, pp. 999–1002, ISBN: 978-1-4673-1961-4. DOI: 10.1109/ISBI.2014.6868041.

[14] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015, ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2014.11.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037014000464.

[15] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, 2015. DOI: https://doi.org/10.3322/caac.21262. eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21262. [Online]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21262.

[16] A. Albayrak and G. Bilgin, "Mitosis detection using convolutional neural network based features," Nov. 2016, pp. 000 335–000 340. DOI: 10.1109/CINTI.2016.7846429.

[17] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 2560–2567. DOI: 10.1109/IJCNN.2016.7727519.

[18] S. Suzuki, X. Zhang, N. Homma, *et al.*, "Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis," in *2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, Tsukuba, Japan: IEEE, Sep. 2016.

[19] I. Wichakam and P. Vateekul, "Combining deep convolutional networks and svms for mass detection on digital mammograms," Feb. 2016, pp. 239–244. DOI: 10.1109/KST.2016.7440527.

[20] H. Xie, J. Li, Z. Qiaosheng, and Y. Wang, "Comparison among dimensionality reduction techniques based on random projection for cancer classification," *Computational Biology and Chemistry*, vol. 65, Sep. 2016. DOI: 10.1016/j.compbiolchem.2016.09.010.

[21] P. Danaee, R. Ghaeini, and D. Hendrix, "A deep learning approach for cancer detection and relevant gene identification," vol. 22, Feb. 2017, pp. 219–229. DOI: 10.1142/9789813207813_0022.

[22] V. Teixeira, R. Camacho, and P. G. Ferreira, "Learning influential genes on cancer gene expression data with stacked denoising autoencoders," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017, pp. 1201–1205. DOI: 10.1109/BIBM.2017.8217828.

[23] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," *Computer Methods and Programs in Biomedicine*, vol. 153, Sep. 2017. DOI: 10.1016/j.cmpb.2017.09.005.

[24] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, and Q. Sun, "Deep learning for image-based cancer detection and diagnosisa survey," *Pattern Recognition*, vol. 83, pp. 134–149, 2018, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2018.05.014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320318301845.

[25] B. Lyu and A. Haque, "Deep learning based tumor type classification using gene expression data," Aug. 2018, pp. 89–96. DOI: 10.1145/3233547.3233588.

[26] S. Hussein, P. Kandel, C. W. Bolan, M. B. Wallace, and U. Bagci, "Lung and pancreatic tumor characterization in the deep learning era: Novel supervised and unsupervised learning approaches," en, *IEEE Trans Med Imaging*, vol. 38, no. 8, pp. 1777–1787, Jan. 2019.

[27] S. Ramroach, A. Joshi, and M. John, "Optimisation of cancer classification by machine learning generates enriched list of candidate drug targets and biomarkers," *Molecular Omics*, vol. 16, Feb. 2020. DOI: 10.1039/C9MO00198K.

[28] S. Sheet, A. Ghosh, R. Ghosh, and A. Chakrabarti, "Identification of cancer mediating biomarkers using stacked denoising autoencoder model - an application on human lung data," *Procedia Computer Science*, vol. 167, pp. 686–695, 2020, International Conference on Computational Intelligence and Data Science, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2020.03.341. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920308073.

[29] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101 693, 2020, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2020.101693. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136184152030058X.

[30] X. Zhang, Y. Xing, K. Sun, and Y. Guo, "Omiembed: A unified multi-task deep learning framework for multi-omics data," *Cancers*, vol. 13, p. 3047, Jun. 2021. DOI: 10.3390/cancers13123047.

[31] R. V.K, N. Arya, S. Ahmad, *et al.*, "Detection of breast cancer using histopathological image classification dataset with deep learning techniques," *BioMed Research International*, vol. 2022, pp. 1–13, Mar. 2022. DOI: 10.1155/2022/8363850.

[32] P. Freitas, F. Silva, J. Sousa, *et al.*, "Machine learning-based approaches for cancer prediction using microbiome data," *Scientific Reports*, vol. 13, Jul. 2023. DOI: 10.1038/s41598-023-38670-0.

[33] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 1, pp. 17–48, 2023. DOI: https://doi.org/10.3322/caac.21763. eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21763. [Online]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21763.