International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

Feature Selection Methods of Gene Expression Based on Machine Learning: A Review

Karwan Jameel Merceedi*, Adnan Mohsin Abdulazeez

Duhok Polytechnic University, Iraq

Email: *karwan.jamil@dpu.edu.krd

Abstract. This article offers a thorough analysis of feature selection strategies that use machine learning to analyze gene expression data. In order to extract significant biological insights, the explosion of high-dimensional genomic data has required the invention and use of sophisticated analysis techniques. In this situation, feature selection is essential because it finds the most pertinent genes that have a major impact on the prediction ability of machine learning models. The paper examines a range of feature selection techniques, classifying them into filter, wrapper, and embedding approaches, each having special advantages and disadvantages. The importance of gene expression data in comprehending the molecular mechanisms underlying complicated diseases and biological processes. The difficulties presented by high-dimensional datasets are next explored, with a focus on feature selection as a means of enhancing model interpretability, lowering computational cost, and raising prediction accuracy. In order to shed light on the fundamental ideas and practical uses of well-known feature selection algorithms, the writers thoroughly examine a number of them, including Mutual Information, Relief, and Recursive Feature Elimination (RFE). Additionally, the study assesses these methods' performance critically across a range of datasets and experimental situations, emphasizing important factors like interpretability, scalability, and resilience. The paper also discusses new developments in feature selection, such as the incorporation of deep learning techniques, ensemble methods, and domain expertise. In order to fully realize the promise of gene expression data for biomedical research and clinical applications, the study ends with a discussion of the present issues and prospective future directions in the field. This discussion emphasizes the significance of creating reliable and understandable feature selection techniques. This thorough study will be an invaluable tool for practitioners, researchers, and bioinformaticians in the field of genomics as they navigate the challenging terrain of feature selection techniques in the context of machine learning-based gene expression analysis.

Keywords: Machine learning, Gene expression, Feature selection.

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

1. Introduction

Within the fields of computational biology and bioinformatics, gene expression data a type of medical data is used to indicate the condition and function of every gene in an organism's genome. This data may be useful in the diagnosis of conditions like cancer (Abdulqader et al., 2020). But out of all the genes in the genome, only a few are relevant when it comes to cancer diagnosis. Therefore, it is crucial to isolate these particular genes from the expression of the complete genome. By using the expression values of a few genes, it is possible to classify a patient as either having the disease or not.

Pattern recognition, statistics, and data mining all include the dynamic field of feature selection as one of their subfields. The fundamental idea behind feature selection is the methodical curation of a subset of input variables, with the purpose of methodically removing those with little to no predictive value. This deliberate method often produces models with better generalization to unknown data points and has the potential to significantly improve the interpretability of the resulting classifier models. Furthermore, the search for the exact subset of predictive traits is inherently important. For example, when making medical decisions, doctors might depend on certain characteristics to assess if expensive surgical procedures are required for the patient's condition (Abdulqader et al., 2020).

In this article, feature selection techniques used in machine learning algorithms for gene expression analysis are thoroughly reviewed. A crucial preprocessing stage that helps identify the most pertinent genes or traits that contribute to a specific biological phenomenon is feature selection (Almazrua & Alshamlan, 2022). The main objective is to separate the signal from the noise so that genetic regulatory networks can be understood more precisely and with greater focus.

The analysis of several feature selection techniques within the machine learning paradigm is the basis of this review. The usefulness of both more recent and sophisticated statistical methods and older established statistical approaches in detecting genes with biological significance and reducing the effects of dimensionality will be examined. We'll pay particular attention to the filter, wrapper, and embedding approaches, as they each provide unique benefits in certain situations (Bommert et al., 2022).

Moreover, the article will illuminate the challenges inherent in gene expression data, such as noise, heterogeneity, and class imbalance, and how feature selection methods aim to address these issues. Real-world applications and case studies will be presented, showcasing successful implementations of machine learning-based feature selection in unraveling the genetic basis of diseases, drug responses, and other biological phenomena.

In conclusion, this review seeks to provide a comprehensive overview of the current landscape of feature selection methods applied to gene expression analysis using machine learning. By synthesizing the existing knowledge, we aim to offer researchers and practitioners a roadmap for navigating the intricate terrain of genomics, fostering a deeper understanding of the intricate dance of genes within the cellular orchestra.

2. Machine Learning

Machine learning is a branch of artificial intelligence (AI) that focuses on developing algorithms and models that enable computers to learn from data and make predictions or decisions without explicit programming. It involves the use of statistical techniques and algorithms to enable computers to improve their performance on a specific task through experience. Machine learning applications are diverse and range from image and speech

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

recognition to recommendation systems, natural language processing, and autonomous vehicles (Srinivasa et al., 2020).

There are three main types of machine learning: supervised learning, where the model is trained on labeled data; unsupervised learning, where the model identifies patterns in unlabeled data; and reinforcement learning, where the model learns by interacting with an environment and receiving feedback. Machine learning plays a crucial role in various industries, driving advancements in healthcare, finance, marketing, and more, by leveraging the power of data to extract meaningful insights and enhance decision-making processes (Srinivasa et al., 2020).

Machine learning, at its core, is a paradigm that empowers computers to learn and adapt without being explicitly programmed. It relies on the utilization of algorithms that enable machines to recognize patterns, make predictions, and improve their performance over time based on experience. The process involves feeding large amounts of data into a model, allowing it to identify underlying patterns and relationships. This learning process is broadly categorized into three types (Tabl et al., 2019a).

Supervised learning involves training a model on a labeled dataset, where the algorithm learns to map input data to corresponding output labels. This type of learning is prevalent in tasks such as image and speech recognition. Unsupervised learning, on the other hand, deals with unlabeled data, and the model's goal is to identify inherent patterns and structures within the information, common applications include clustering, anomaly detection, and dimensionality reduction (Jo, 2021). Reinforcement learning, the third type, involves an agent learning to make decisions by interacting with an environment and receiving feedback in the form of rewards or penalties. This approach is particularly relevant in scenarios like autonomous systems and game playing (Jo, 2021).

The impact of machine learning is profound, permeating various industries and reshaping the way we approach problem-solving. In healthcare, it aids in disease diagnosis and personalized treatment plans. Financial institutions leverage machine learning for fraud detection and risk assessment, while marketing benefits from personalized recommendations and targeted advertising. As technology continues to advance, machine learning's role is expected to expand further, driving innovation and enhancing efficiency across diverse domains. Machine learning can be broadly categorized into three main types, each with its unique approach to learning from data (Al-Azzam & Shatnawi, 2021) (Abdulqader et al., 2020).

In supervised learning, the algorithm is trained on a labeled dataset, where the input data is paired with corresponding output labels (Lindholm et al., n.d.). The goal is for the model to learn the mapping or relationship between the input and output so that it can make predictions or decisions when presented with new, unseen data. Common applications include classification, where the algorithm predicts the class or category of an input, and regression, where it predicts a continuous value (Jiang et al., 2020) (Sen et al., 2020).

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

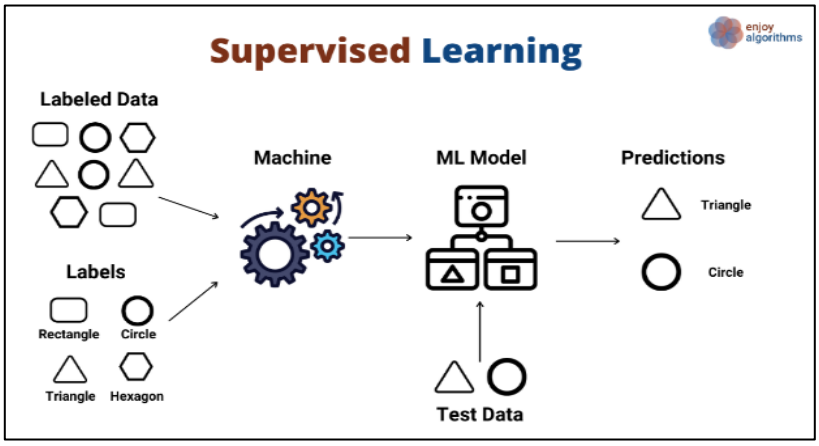

Figure 1. Supervised Learning Model

Unsupervised learning deals with unlabeled data, and the algorithm's objective is to identify patterns, structures, or relationships within the data without explicit guidance. Clustering is a common unsupervised learning task, where the algorithm groups similar data points into clusters (Scheurer & Slager, 2020). Another task is dimensionality reduction, which involves simplifying the data while retaining its essential features. Unsupervised learning is valuable for exploring and discovering hidden patterns in data without predefined categories (Usama et al., 2019) (Ceriotti, 2019).

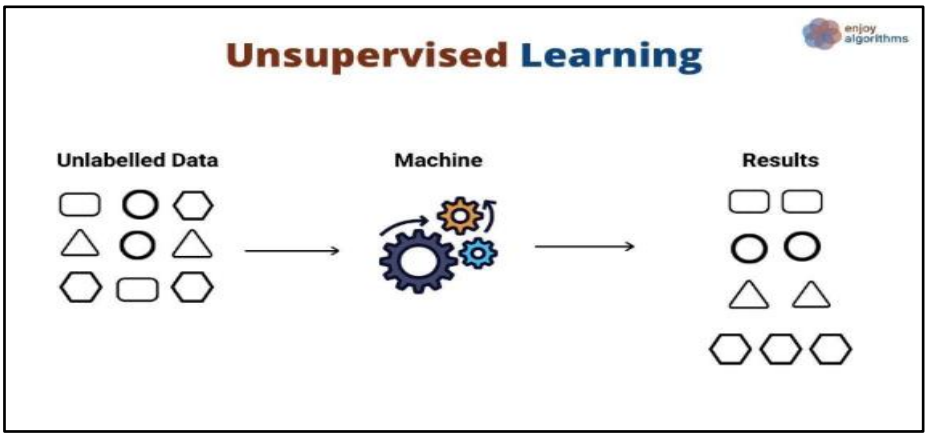

Figure 2. Unsupervised Learning Model

Reinforcement learning involves an agent interacting with an environment and learning to make decisions by receiving feedback in the form of rewards or penalties. The agent aims to learn a policy, a set of actions, that maximizes the cumulative reward over time. This type of learning is prevalent in applications such as game playing, robotics, and autonomous systems, where an agent learns to navigate and adapt to its surroundings through trial and error (Al-Azzam & Shatnawi, 2021) (Berry et al., 2020).

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

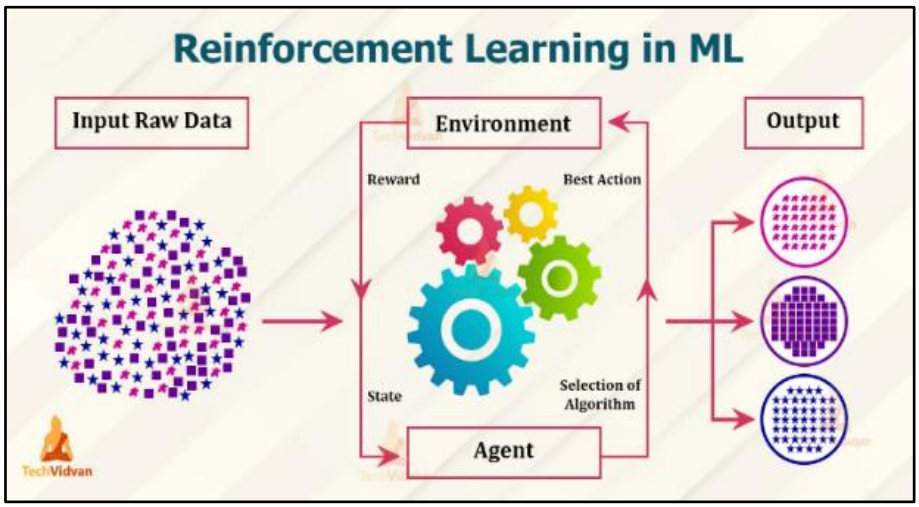

Figure 3. Reinforcement Learning Model

These three types of machine learning represent different approaches to solving problems and extracting insights from data, and often, a combination of these techniques is employed for more complex tasks.

2.1. Machine Learning Algorithms

Machine learning algorithms are the backbone of artificial intelligence systems, enabling computers to learn and make predictions or decisions without explicit programming. These algorithms are designed to identify patterns, learn from data, and improve their performance over time. There are various types of machine learning algorithms, broadly categorized into supervised learning, unsupervised learning, and reinforcement learning (Ray, n.d.).

Trained on labelled datasets, where the input data is paired with corresponding output labels. The algorithm learns to map inputs to outputs, making predictions on new, unseen data. Common supervised learning algorithms include linear regression, decision trees, support vector machines, and neural networks (Uddin et al., 2019).

Unsupervised learning, on the other hand, deals with unlabelled data. The algorithm explores the inherent structure or patterns within the data without explicit guidance. Clustering algorithms, such as k-means and hierarchical clustering, fall under unsupervised learning, as do dimensionality reduction techniques like principal component analysis (PCA) and autoencoders (Mahesh, 2018).

Reinforcement learning involves training algorithms to make sequential decisions by interacting with an environment. The algorithm receives feedback in the form of rewards or penalties, adjusting its actions to maximize cumulative rewards over time. This approach is commonly used in applications like game playing, robotic control, and autonomous systems (Ferdous et al., 2020).

Furthermore, within these broad categories, there are numerous specialized algorithms tailored for specific tasks. Random forests, gradient boosting, and ensemble methods are popular techniques for improving predictive performance (Ngiam & Khor, 2019). Support vector machines excel in classification tasks, while recurrent neural networks and long short-term memory networks are well-suited for sequential data, like time series or natural language (O. Ahmed & Brifcani, 2019).

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

Machine learning algorithms are versatile and find applications in various domains, including healthcare, finance, image and speech recognition, natural language processing, and recommendation systems. The success of these algorithms depends on the quality and quantity of data, as well as careful consideration of the problem at hand when selecting the most suitable algorithm. As technology continues to advance, machine learning algorithms play a crucial role in shaping the capabilities of intelligent systems and driving innovation across industries (Sarker, 2021).

2.1.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful and widely used supervised machine learning algorithm that falls under the category of discriminative classifiers. Developed by Vladimir Vapnik and his colleagues in the 1990s, SVM is particularly effective for classification and regression tasks. The primary objective of SVM is to find a hyperplane that best separates data points into different classes while maximizing the margin between the classes (Sarker, 2021).

The key idea behind SVM is to identify the optimal decision boundary (hyperplane) that maximally separates data points of different classes. The optimal hyperplane is the one that has the maximum margin, defined as the distance between the hyperplane and the nearest data points of each class. SVM aims to find this hyperplane in a high-dimensional space, where each feature of the input data corresponds to a dimension (Kang et al., 2019a).

One of the strengths of SVM is its ability to handle both linear and non-linear classification tasks. In a linearly separable case, where the classes can be separated by a straight line, SVM determines the optimal hyperplane using methods like the maximal margin hyperplane or the support vector approach. For non-linearly separable cases, SVM can use kernel functions, such as polynomial or radial basis function (RBF) kernels, to map the input data into a higher-dimensional space where a hyperplane can effectively separate the classes (Berry et al., 2020).

The term "support vectors" in SVM refers to the data points that are crucial in defining the optimal hyperplane. These are the data points that lie closest to the decision boundary and influence the positioning and orientation of the hyperplane. SVM relies on these support vectors to make predictions for new, unseen data (Lindholm et al., n.d.).

SVM has proven to be effective in various applications, including text classification, image recognition, bioinformatics, and finance. Its robust performance is attributed to its ability to handle high-dimensional data, its resistance to overfitting, and its versatility in handling both linear and non-linear relationships within the data (Burkart & Huber, 2021).

Despite its strengths, SVM's performance may be affected by the choice of kernel function and its sensitivity to the parameters. Additionally, SVM might face challenges when dealing with large datasets, as the computational complexity increases with the number of data points (Al-Azzam & Shatnawi, 2021).

In summary, Support Vector Machine is a versatile and powerful algorithm that has found success in various machine learning applications due to its ability to handle different types of data and perform well in both linear and non-linear scenarios (Usama et al., 2019).

2.1.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple yet powerful supervised machine learning algorithm used for classification and regression tasks. It falls under the category of instance-based or lazy learning methods, as it does not explicitly build a model during the training phase. Instead, KNN stores the entire training dataset and makes predictions by comparing the input data with the stored instances (Shaban et al., 2020a).

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

The fundamental principle behind KNN is based on the assumption that similar data points in the feature space tend to belong to the same class or exhibit similar behavior. The "K" in KNN refers to the number of nearest neighbors that influence the prediction for a given data point. When making a prediction for a new data point, the algorithm identifies the K nearest neighbors in the training dataset based on a chosen distance metric, such as Euclidean distance or Manhattan distance (Sun et al., 2019a).

For classification tasks, KNN takes a majority vote among its K nearest neighbors to assign a class label to the new data point. In regression tasks, the algorithm computes the average (or weighted average) of the target values of the K nearest neighbors to predict a continuous output (Taunk et al., 2019).

One of the strengths of KNN is its simplicity and ease of implementation. Additionally, KNN can adapt to different types of data and does not make strong assumptions about the underlying distribution of the data. However, the performance of KNN can be sensitive to the choice of the distance metric and the value of K. Selecting an appropriate value for K is crucial; a small K can make the algorithm sensitive to noise, while a large K may lead to oversmoothing (Xing & Bei, 2020).

Despite its simplicity, KNN has proven effective in various applications, including image recognition, handwritten digit classification, and recommendation systems. However, its computational complexity can be a limitation, particularly with large datasets, as the algorithm needs to compute distances for each new instance with respect to all training instances during prediction (Bansal et al., 2022).

In summary, K-Nearest Neighbors is a versatile and intuitive algorithm that leverages the concept of proximity in feature space to make predictions. While it may not be suitable for all scenarios, it remains a valuable tool in the machine learning toolkit, especially for small to moderately sized datasets and when interpretability is crucial (Shokrzade et al., 2021).

2.1.3 K-Means

K-Means is a widely used clustering algorithm in machine learning and data analysis. It falls under the category of unsupervised learning algorithms, specifically designed for partitioning a dataset into K distinct, non-overlapping subsets or clusters. The goal of K-Means is to group similar data points together and assign them to clusters, where the number of clusters (K) is a user-defined parameter (Sinaga & Yang, 2020).

The algorithm operates iteratively, starting with an initial set of K cluster centroids randomly placed in the data space. It then alternates between two steps until convergence. In the assignment step, each data point is assigned to the cluster whose centroid is closest, typically measured using Euclidean distance. In the update step, the centroids of the clusters are recalculated based on the mean of the data points within each cluster (M. Ahmed et al., 2020).

One challenge in using K-Means is that the algorithm's performance can be sensitive to the initial placement of centroids. To mitigate this, K-Means often employs multiple random initializations, and the result with the lowest overall intra-cluster variance (sum of squared distances from data points to their assigned cluster centroids) is chosen (Hassan et al., 2021).

K-Means has found applications in various fields, including image segmentation, customer segmentation in marketing, anomaly detection, and document clustering. Its simplicity and efficiency make it suitable for large datasets and situations where the underlying data distribution is relatively well-behaved (X. Liu et al., 2018).

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

However, K-Means has limitations. It assumes that clusters are spherical and equally sized, which may not be appropriate for all types of data. The algorithm is also sensitive to outliers, and the choice of the number of clusters (K) can impact the results significantly (M. Ahmed et al., 2020).

Researchers and practitioners often use variations of K-Means, such as K-Means++, which improves the initialization step to enhance convergence speed and reduce sensitivity to initialization. Despite its limitations, K-Means remains a fundamental and widely applied clustering algorithm, providing valuable insights into the structure of unlabeled datasets (C. Yuan & Yang, 2019).

2.1.3 Naïve Bayes

simple probabilistic classification algorithm based on Bayes' theorem, which is a fundamental probability theorem. Despite its simplicity, Naive Bayes often performs surprisingly well in various real-world applications, particularly in natural language processing tasks like spam filtering and text classification (Chen et al. - 2020 - A Novel Selective Naïve Bayes Algorithm.Pdf, n.d.).

The "naive" in Naive Bayes comes from the assumption of independence among features. This means that the algorithm assumes that the presence of a particular feature in a class is independent of the presence of other features. Although this assumption might not hold true in all cases, it simplifies the computation and allows the algorithm to be computationally efficient and easy to implement (Chen et al., 2020).

The algorithm works by calculating the probability of a given instance belonging to a particular class based on the features observed in that instance. Bayes' theorem is used to update these probabilities as new features are considered (Surya and Subbulakshmi - 2019 - Sentimental Analysis Using Naive Bayes Classifier.Pdf, n.d.). The formula for Bayes' theorem is:

$P(\text{class}|\text{features}) = (P(\text{features}|\text{class}) \times P(\text{class})) / P(\text{features})$ In this formula:

- $P(\text{class} \mid \text{features})$ is the probability of the class given the observed features.
- $P(\text{features} \mid \text{class})$ is the probability of the features given the class.
- $P(\text{class})$ is the prior probability of the class.
- $P(\text{features})$ is the probability of the observed features.

The algorithm classifies an instance by selecting the class with the highest posterior probability. Despite its simplifying assumptions, Naive Bayes often performs surprisingly well and can be competitive with more complex algorithms, especially when dealing with high-dimensional data.

There are different variants of Naive Bayes, such as Gaussian Naive Bayes (for continuous data assuming a Gaussian distribution), Multinomial Naive Bayes (for discrete data, commonly used in text classification), and Bernoulli Naive Bayes (for binary data) (Itoo et al., 2021).

Overall, Naive Bayes is a powerful and efficient algorithm, particularly suitable for tasks where the assumption of feature independence is reasonable. Its simplicity, speed, and effectiveness make it a popular choice for a wide range of applications (Rahat et al., 2019).

3. Feature Section

Feature selection is a crucial step in the process of machine learning and data analysis, aimed at identifying and retaining the most relevant variables or features from a given dataset. The primary objective is to enhance model performance by reducing dimensionality,

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

mitigating the risk of overfitting, and improving interpretability (Ali & Aittokallio, 2019). In essence, feature selection involves choosing a subset of features that contributes the most to the predictive power of a model while discarding redundant or irrelevant information. Various techniques are employed for feature selection, ranging from filter methods that assess feature relevance independently of the chosen learning algorithm, wrapper methods that incorporate the predictive model's performance, to embedded methods where feature selection is an integral part of the model training process (Toğaçar et al., 2020). The benefits of effective feature selection are manifold, including faster model training, enhanced generalization, and a clearer understanding of the underlying patterns within the data. However, the choice of an appropriate feature selection method depends on factors such as the nature of the data, the learning algorithm, and the specific objectives of the analysis or model (P. Ghosh et al., 2021a).

Feature selection is a critical aspect of the model-building process, influencing the model's performance, efficiency, and interpretability. In essence, the goal is to identify and retain a subset of features that significantly contribute to the predictive power of the model while eliminating irrelevant or redundant variables. The need for feature selection arises from the curse of dimensionality, where an excessive number of features relative to the number of observations can lead to increased computational complexity, decreased model interpretability, and a higher risk of overfitting. There are three main categories of supervised feature selection methods: filter methods, wrapper methods, and embedded methods (Sun et al., 2019a).

3.1 Supervised Feature Selection Techniques

3.1.1 Filter Methods:

These methods assess the relevance of features independent of any specific machine learning algorithm. Common techniques include statistical tests, correlation analysis, and information gain. Features are ranked or scored based on their individual characteristics, and a threshold is applied to select the most informative ones (Kang et al., 2019a).

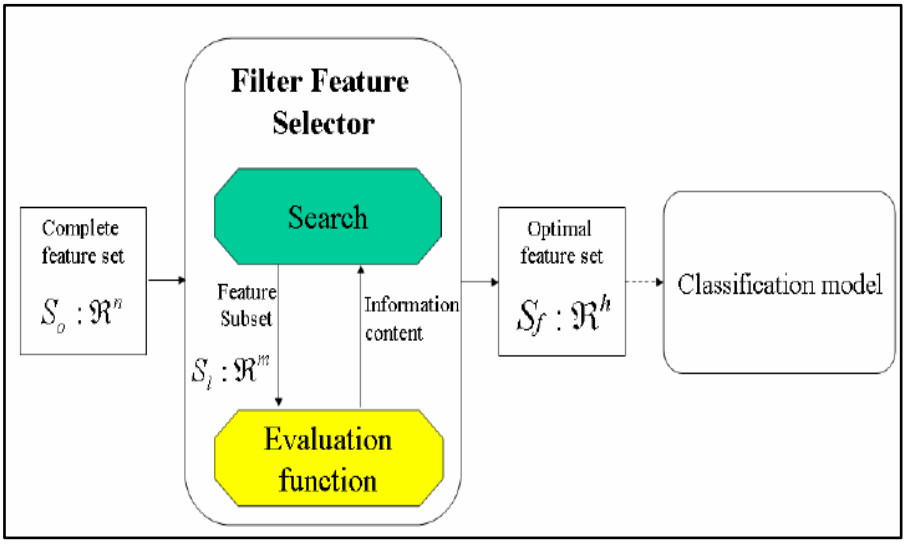

Figure 4. Feature Selection Methods: Filter Method

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

3.1.2 Wrapper Methods:

Unlike filter methods, wrapper methods incorporate the performance of the machine learning model during the feature selection process. They involve repeatedly training and evaluating the model with different subsets of features to identify the optimal set. Examples include forward selection, backward elimination, and recursive feature elimination (RFE) (M. Ghosh et al., 2020a).

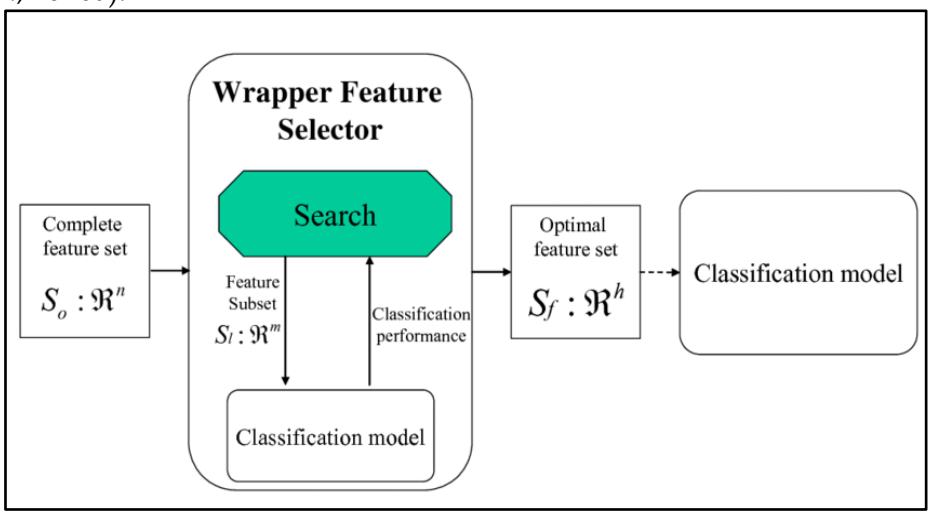

Figure 5. Feature Selection Methods: Wrapper Method

3.1.3 Embedded Methods:

These methods integrate feature selection into the model training process itself. Certain machine learning algorithms, such as decision trees and regularization-based models like LASSO (Least Absolute Shrinkage and Selection Operator), inherently perform feature selection as part of their optimization process. This integration often results in more efficient models by directly penalizing or pruning less informative features (H. Liu et al., 2019) (Almugren & Alshamlan, 2019).

The benefits of effective feature selection are multifaceted. It not only reduces the computational burden by working with a subset of relevant features but also improves model generalization to new, unseen data. Additionally, a concise set of features enhances model interpretability, allowing stakeholders to better understand the driving factors behind predictions. However, the choice of a specific feature selection method depends on various factors, including the nature of the data (categorical, numerical, or mixed), the characteristics of the features (linearly correlated or nonlinear), the size of the dataset, and the specific goals of the analysis. Striking a balance between model simplicity and predictive accuracy is essential, and practitioners often iterate through different feature selection techniques to optimize their models for a given task (H. Liu et al., 2019).

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

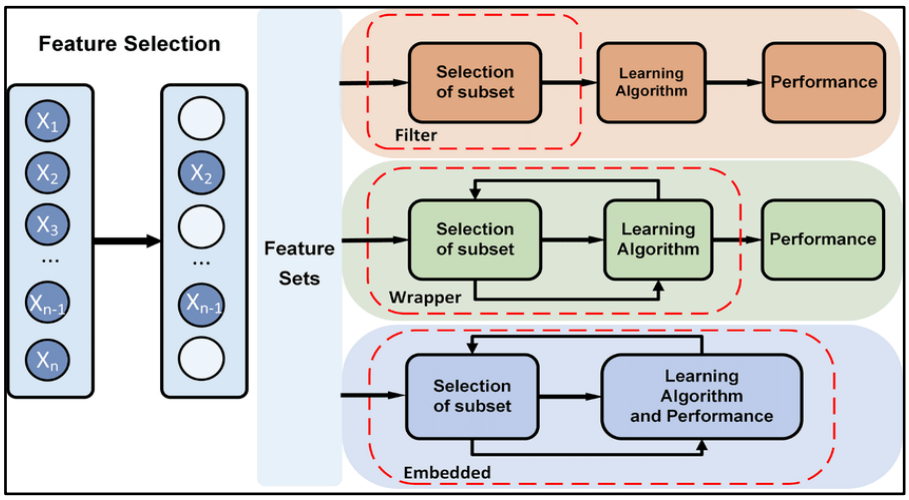

Figure 6. Feature Selection Methods: Embedded Method

3.2 Unsupervised Feature Selection Techniques

These are techniques where you're using an algorithm to find patterns and similarities in data without explicit instructions. In other words, without telling the algorithm what's good and what's not, i.e., features are selected without reference to a target variable (Solorio-Fernández et al., 2020).

Without requiring labeled data, the techniques let you investigate and identify significant features in the data. It's like giving the computer a puzzle and letting it figure out connections on its own with these machine learning feature selection algorithms. Without your assistance, they will arrange data and find patterns. But we'll discuss three unsupervised feature selection methods for machine learning in this article (Solorio-Fernández et al., 2020).

3.2.1 Principal Component Analysis (PCA)

PCA is a technique for deciphering and organizing data. It assists us in identifying the key components of the data. Consider that you have a large, intricate picture. PCA assists us in determining the primary forms or colors that are most noticeable. It's similar to identifying the main components that sum up the scene without getting bogged down in the minutiae (Haq et al., 2021).

3.2.2 Independent Component Analysis (ICA)

ICA is a feature selection strategy that aids in our comprehension of how various elements come together. Picture yourself holding a box full of disparate noises, such as music playing, people conversing, and cars honking. ICA will assist us in distinguishing between those sounds and identifying each sound on its own. In order to comprehend what each person, thing, or voice is saying or performing, it's similar to paying close attention and differentiating between the voices or instruments in a crowded area (Tharwat, 2021).

3.2.3 Non-negative Matrix Factorization (NMF)

Using the NMF approach, we may decompose large numbers into smaller positive numbers. For example, let's say you have a large number that represents an entire picture, and you want to know what components make up that picture. NMF assists us in identifying tiny positive numbers that, when added together, reproduce the larger number or image. It's similar to disassembling a puzzle and learning how the tiny parts fit together to form the larger image (A. Yuan et al., 2022).

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

4. Gene Expression

Gene expression is a fundamental process in molecular biology that describes how information encoded in a gene is utilized to synthesize functional gene products, primarily proteins. This intricate process involves the transcription of a gene's DNA sequence into a complementary RNA molecule, known as messenger RNA (mRNA), in a cellular structure called the nucleus. The synthesis of mRNA is mediated by RNA polymerase, which reads the DNA template and assembles the corresponding RNA sequence. Following transcription, the mRNA exits the nucleus and enters the cytoplasm, where it serves as a template for protein synthesis during translation (Pinal-Fernandez et al., 2020).

Gene expression is a tightly regulated and highly dynamic process that allows cells to respond to environmental cues, developmental signals, and various physiological demands. The regulation occurs at multiple levels, including transcriptional, post-transcriptional, translational, and post-translational mechanisms. Transcriptional regulation involves the control of RNA polymerase activity and the accessibility of the DNA template through the binding of transcription factors to specific regulatory regions (Seal et al., 2020a).

Moreover, post-transcriptional processes, such as alternative splicing and RNA modification, contribute to the diversity of mRNA isoforms and impact the final protein product. Translation, the subsequent step, involves the conversion of mRNA into a functional protein with the help of ribosomes and transfer RNA (tRNA). Finally, post-translational modifications, such as phosphorylation, acetylation, and glycosylation, play a crucial role in modulating protein function, stability, and localization (Khalifa et al., 2020a).

Dysregulation of gene expression can lead to various diseases, including cancer, neurodegenerative disorders, and metabolic conditions. Researchers aim to decipher the complexities of gene expression to better understand cellular functions, disease mechanisms, and potential therapeutic targets. Advanced technologies, such as RNA sequencing and CRISPR-Cas9 gene editing, have significantly contributed to our ability to study and manipulate gene expression, opening new avenues for precision medicine and biotechnological advancements. In summary, gene expression is a central and dynamic process essential for the proper functioning and adaptation of cells in diverse biological contexts (Abdulqader et al., 2020).

Gene selection, also known as feature selection in the context of machine learning and bioinformatics, is a crucial step in the analysis of high-dimensional biological data, particularly in the field of genomics. Genes are segments of DNA that encode information for the synthesis of proteins and play a fundamental role in determining an organism's traits and functions. However, not all genes are relevant or contribute significantly to a specific biological process or disease (Almugren & Alshamlan, 2019).

The objective of gene selection is to identify a subset of genes from the vast pool of available genes that are most informative for a particular task, such as classifying different disease states or understanding the underlying mechanisms of a biological phenomenon. This process is essential for reducing the dimensionality of data and improving the efficiency and interpretability of subsequent analyses (Kegerreis et al., 2019).

Several methods are employed for gene selection, ranging from statistical techniques to machine learning algorithms. Statistical approaches often involve measures such as t-tests, ANOVA, or correlation coefficients to assess the significance of individual genes in relation to a specific outcome. Machine learning-based methods, on the other hand, leverage algorithms like decision trees, support vector machines, or feature importance scores from models like random forests (Maniruzzaman et al., 2019).

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

Gene selection has wide-ranging applications, including the identification of biomarkers for diseases, understanding the genetic basis of complex traits, and improving the accuracy of predictive models. The significance of gene selection becomes particularly evident in scenarios where the number of features (genes) is much larger than the number of samples, a common challenge in genomics data (F. Yuan et al., 2020).

In summary, gene selection is a critical step in the analysis of genomic data, playing a pivotal role in enhancing the understanding of biological processes, identifying potential therapeutic targets, and facilitating the development of more accurate diagnostic and predictive models in the realm of personalized medicine (Hossain et al., 2019).

5. Literature Survey

The authors in (Mohammed & Abdulazeez, 2017) proposed a Mahalanobis distance, Partition around medoids (PAM) algorithm, Weighted features of Microarray expression datasets, Dunn's validity index, and Weighted Normalised Mahalanobis distance techniques, with Microarray datasets were used in the study to obtain the results as Optimal cluster solution for Hypoxia dataset when k=3, Dunn's index value of 0.0783 for both Hypoxia and ATMs datasets, improved cluster quality with proposed algorithm using Weighted and Normalized Mahalanobis distance, highest Dunn's index value for ATMs dataset when k=2, enhanced performance of PAM algorithm with proposed distance on ATMs dataset, therefore, identification of relevant gene patterns in microarray data, proposal of an enhanced PAM algorithm based on weighted Normalised Mahalanobis distance, improvement of cluster quality in microarray expression data using the proposed algorithm.

While authors in (Zeebaree et al., 2018) proposed a Multilayered CNN (Convolutional Neural Network) algorithm, Deep learning algorithm with integration of strongly linked cancer datasets, and detection of latent characteristics of cancer from comparable types through in ten cancer datasets are used in the study. However, they attained an ANOVA analysis showed statistical significant difference between the three methods. Proposed CNN had a mean classification accuracy of 94.74, best accuracy performance of proposed CNN was 100, mSVM-RFE-IRF had a mean classification accuracy of 85.82, best accuracy performance of mSVM-RFE-IRF was 95.55, varSelRF had a mean classification accuracy of 79.58, best accuracy performance of varSelRF was 93.07, proposed CNN had the highest accuracy in Brain dataset (92.14), proposed CNN had the highest accuracy in Breast3 dataset (92.90), proposed CNN had the highest accuracy in Leukemia dataset (95.69), and proposed CNN had the highest accuracy in Lymphoma dataset (100.00). Proposed CNN scored higher accuracies compared to other methods in cancer datasets, K-NN classifier performed better than random forest in accurate classification, Random Survival Forest strategy for selecting informative genes, PSOC4.5 hybrid used for classifying informative genes with superior accuracy.

The author in (H. Al-Baity & Al-Mutlaq, 2021) suggested a novel, enhanced technique for choosing wrapper genes, called simulated annealing (SA), which draws inspiration from nature and aids in identifying the most informative genes for breast cancer prognosis. The decision tree, random forest, and SVM were the three supervised machine-learning algorithms that were utilized to develop the classifier models that will aid in breast cancer prediction. Three datasets gene expression (GE), deoxyribonucleic acid (DNA) methylation and a mixture of the two were used in two distinct research. The outcomes showed that, this method performed better than traditional classifiers. SA-SVM has produced high accuracy values of 99.77%, 99.45%, and 99.45% for the combined dataset, GE, and DNA methylation, respectively. The suggested method's execution time was much shortened, the SA-SVM achieved the best

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

execution time which is 0.02, 0.03, and 0.02 on the GE, DNA methylation, and combined datasets.

On the other hand, authors in (Kang et al., 2019b) suggested a novel approach to tumor categorization called relaxed Lasso-GenSVM (rL-GenSVM). The tumor dataset is first standardized using z-score and split into training and test sets. Second, on the training set relaxed Lasso chooses feature genes. GenSVM functions as the classifier in the process of determining the ideal parameters using a 10-fold cross-validation grid search on the training set. The outcomes of the experiment demonstrate that the suggested approach choose fewer feature genes while achieving a greater classification accuracy. However, Regularization parameters are used by (rL-GenSVM) to prevent overfitting, and it is generally applicable to the classification of high-dimensional and small-sample tumor data.

While authors in (Khan et al., 2019) proposed a two-stage gene selection strategy that finds the most discriminative genes. In the first step, genes that clearly classify the maximum number of samples into each class using a greedy method are chosen. There are a specific number of clusters made up of the remaining genes. The lasso approach is used to choose the most informative genes from each cluster, which are then integrated with the genes chosen in the first step. In order to accomplish this, two classifiers the random forest and support vector machine that are applied to datasets containing specific genes and training samples. The result show that when compared to other techniques, the GClust method has better results and has the highest accuracy.

Then in (Khorshid & Abdulazeez, 2021) the authors are proposed a machine learning algorithms like K-NN, Computer-Aided Diagnosis (CAD), Artificial Neural Network (ANN), Support Vector Machine (SVM), and Principal Component Analysis (PCA) in own study with used a Mammograms from the Automated Mammography Screening Database (DDSM), breast mammograms from Mini MIAS (Mammographic Image Processing Society), Wisconsin Breast Cancer Diagnosis (WDBC) data set from UCI machine learning repository datasets to get the results as SVM has the highest accuracy of 98% in image processing techniques, LR and LDA have accuracies of 97.23% and 95.73% respectively, K-NN technique achieved the best results compared to NB and CART, SVM has the highest accuracy of 97.07% among 8 ML algorithms, ANNs have achieved the best accuracy, precision, and F1 score, SVM and RF Classifier are the best predictive analyzes with an accuracy of 96.5%, ANN and CNN have the highest accuracies of 99.3% and 97.3% respectively, SVM cubic classifier has an accuracy of 92.3%, SVM has an accuracy of 96% and K-NN has an accuracy of 100% in breast cancer detection. Machine learning and mechanistic modeling for prediction of metastatic relapse in early-stage breast cancer, classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer, also breast cancer detection using automated whole breast ultrasound, and pectoral muscle segmentation for cancer detection and diagnosis.

Moreover, the authors in (Cho et al., 2022) proposed the development of a bio-signature of immunotherapy based responses using gene expression data. ML algorithms, such as random forests, deep neural networks (DNN), support vector machines (SVM), along with boosting and feature selection techniques, are effective in classifying immune phenotypes of BC with gene expressions and identifying associations between specific gene expressions and the phenotypes. In order to identify gene expression features useful for immune phenotype classification, the results show that DNN yielded the highest area under the curve (AUC) with precision and recall (PR) curves and receiver operating characteristic (ROC) curves for each phenotype (0.711 ± 0.092 and 0.86 ± 0.039, respectively).

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

In another hand, (Singh et al., 2020) authors suggested the feature selection network (FsNet), a DNN-based, nonlinear feature selection technique for high-dimensional, sparse sample data. FsNet is made up of two specific layers, a reconstruction layer that stabilizes the training and a selection layer that chooses features. Since overfitting can easily occur when there are too many parameters in the selection and reconstruction layers for a small number of samples. The result show that for a high-dimensional data, FsNet can achieve superior performance with a significantly smaller number of parameters.

However, the FSDNE algorithm, which use neighborhood rough sets and neighborhood entropy-based uncertainty measures, is a unique feature selection method that the author proposed in (Sun et al., 2019b). It uses the KNN, C4.5, and SVM classifiers for gene selection and classification for cancer classification. In high-dimensional gene expression datasets, it enhances classification performance by fusing the neighborhood rough sets with the Fisher score. The FSDNE algorithm produced the highest average classification accuracy of 84% using the KNN, followed by C4.5 and SVM classifiers with lower accuracy. The results demonstrate that the proposed method outperforms other related methods in terms of the number of selected genes and classification accuracy.

Where in (Wu & Hicks, 2021) authors proposed the use of gene expression data to classify breast cancer using machine learning algorithms. The effectiveness of machine learning algorithms in utilizing gene expression data to classify breast cancer into non-triple-negative and triple-negative subtypes. The algorithm that performed the best was the support vector machine (SVM), which had 90% accuracy, 87% recall, and 90% specificity. After evaluating four distinct classification models, they discovered that the SVM algorithm outperformed other ML algorithms in terms of accuracy, having higher sensitivity, specificity, and fewer misclassification errors.

Then researchers in (Zhang et al., 2021) proposed a fusion feature selection framework attributed to an ensemble method called Fisher score and Gradient Boosting Decision Tree (FS–GBDT), in order to choose reliable and significant feature genes in high-dimensional gene expression datasets by using machine learning algorithm. To investigate the key feature genes subset of cancer, a collaborative analysis of 11 human cancer types was carried out. In order to confirm the effectiveness of FS-GBDT, a Support Vector Machine (SVM) classifier was used to compare it with four other popular feature selection algorithms. The FS-BDT algorithm using SVM outperforms the other four methods and achieves the highest indicators.

While in (M. Ghosh et al., 2020b) researchers suggested using a wrapper-filter combination of ACO. The practice to construct embedded systems by combining a filter approach with a wrapper method, where the computational complexity is reduced by performing a subset evaluation via a filter method as opposed to a wrapper method. The suggested technique has been tested with K-nearest neighbours and multi-layer perceptron classifiers on a variety of real-world datasets from the UCI Machine Learning repository and the NIPS2003 FS challenge. The results are contrasted with a few well-known FS techniques. The comparison of the findings demonstrates unequivocally that method performs better than the majority of the cutting-edge FS algorithms.

Moreover, the authors in (Tabl et al., 2019b) developed a hierarchical machine learning system that forecasts patients who received a certain therapys 5-year survival. They used machine learning algorithms like Bayesians Naive Bayes, SVM, and Random Forest at each node, the model classifies one class against the others, resulting in the creation of five nodes in the tree-based model. The model uses a hierarchical model as a tree that comprises one-versus-

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

rest classifications, and the findings demonstrate that the model can identify the classes with high-performance measurements and very high accuracy levels.

While in (Khalifa et al., 2020b), Authors suggested a novel optimized deep learning method to classify different forms of cancer based on tumor RNA sequence (RNA-Seq) gene expression data. The method is based on binary particle swarm optimization with decision tree (BPSO-DT) and convolutional neural network (CNN). The three stages of the suggested strategy are as follows. Pre-processing is the initial step, wherein the high-dimensional RNA-seq is first optimized using BPSO-DT to choose only the most important features, and the optimized RNA-seq is subsequently converted to 2D pictures. The second stage, known as augmentation, makes the 2086 sample original dataset five times larger. This stage trains the model to attain higher accuracy. Deep CNN architecture represents the third stage, during this stage, a two-main convolutional layer architecture is used to extract features and categorize the five different forms of cancer. Recall, precision, and F1 score, among other performance indicators, indicate that the suggested strategy produced an overall testing accuracy of 96.90%.

However, in another paper they presented a BukaGini algorithm like in (Bouke et al., 2023), that an innovative and reliable method that takes advantage of the Gini impurity index for feature interaction analysis. The suggested technique successfully captures both linear and nonlinear feature interactions by taking advantage of the special qualities of the BukaGini index, giving the underlying data a richer and more thorough representation. The experimental findings show that the BukaGini algorithm routinely achieves higher accuracy than conventional Gini index-based techniques, the BukaGini algorithm exhibits improvements ranging from 0.32% to 2.50% in all examined datasets, demonstrating its efficacy in managing a wide range of data types and problem domains.

Then in (Kurniabudi et al., 2020) the authors proposed using important and significant elements of massive network traffic to reduce the execution time and increase the accuracy of traffic anomaly identification. The most popular feature selection method in intrusion detection system (IDS) research is information gain. Using the CICIDS-2017 dataset, they conduct trials using the Random Forest, Bayes Net, Random Tree, Naive Bayes and J48 classification methods. The experiment's findings demonstrate that improvements in detection accuracy and execution time are highly correlated with the quantity of pertinent and noteworthy features that Information Gain produces, however, with 22 relevant selected features, the Random Forest algorithm achieves the best accuracy of 99.86%, while the J48 classifier method employs 52 relevant selected features but requires a longer execution time to get an accuracy of 99.87%.

Moreover, the scientists in (P. Ghosh et al., 2021b) presented a model that effectively predicts cardiac disease by combining various techniques. The Relief and Least Absolute Shrinkage and Selection Operator (LASSO) approaches are used to choose appropriate features. By combining the traditional classifiers with bagging and boosting techniques, which are applied during training, new hybrid classifiers are created, such as Decision Tree Bagging Method (DTBM), Random Forest Bagging Method (RFBM), K-Nearest Neighbors Bagging Method (KNNBM), AdaBoost Boosting Method (ABBM) and Gradient Boosting Boosting Method (GBBM). To facilitate comparisons, the outcomes are displayed individually. The suggested model yielded the best accuracy (99.05%) when employing the RFBM and Relief feature selection techniques.

On the other hand, a new virus has emerged called COVID-19, that needs a new algorithm to detect it, the authors in (Shaban et al., 2020b) suggested an algorithm called COVID-19 Patients Detection Strategy (CPDS). There are two main contributions that comprise the

originality of CPDS. The first is a new hybrid feature selection methodology called HFSM, which chooses the important characteristics for the subsequent detection phase. The second addition is an improved K-Nearest Neighbor (EKNN) classifier, which adds reliable heuristics to select the neighbors of the tested item. Consequently, EKNN, using those important features chosen by the HFSM approach, can precisely identify infected patients with the least amount of time penalty. The suggested detection strategy beats more contemporary methods since it introduces the greatest accuracy rate, according to the experimental data.

Furthermore, the authors in (Xia et al., 2021), presented an algorithm that used a supervised method based on a Random Forest that is quick and repeatable. in order to identify significant features from three microarray datasets from prenatal nicotine, alcohol, and nicotine and alcohol exposure groups in two different cell types. This method of reducing the dimensionality of incredibly huge microarray datasets proved computationally efficient. Subsequently, the outcomes demonstrated that utilizing the highest 20% of characteristics was adequate to validate the genetic pathways that had been previously discovered while utilizing every feature in the model.

However, the authors in (Seal et al., 2020b), created a deep learning-based predictive model that can quantitatively capture the relationship between directionality of gene expression for liver hepatocellular carcinoma (LIHC) and genetic and epigenetic changes, by using Deep Denoising Auto-encoder (DDAE) and Multi-layer Perceptron (MLP). The machine learning algorithm that has been trained to identify important characteristics from the input omics data and estimate gene expression uses the DDAE. The findings demonstrate that a deep learning-based integration model has been assessed for its capacity to classify diseases, with a 95.1% accuracy rate.

While a classification technique was proposed by researchers in (Mallick et al., 2023), to comprehend the convergence of deep neural network (DNN) training. Since the network is over-parameterized and the inputs do not degenerate, the assumptions are made. Also, there are a sufficient number of hidden neurons. The authors of this work classified the gene expression data using DNN. Seventy-two leukemia patients' bone marrow expressions are included in the dataset used in this analysis. The classification of acute lymphocyte (ALL) and acute myelocytic (AML) samples is accomplished by a five-layer DNN classifier. 80% of the data is used to train the network, and the remaining 20% is used for validation. The result show that 98.2% accuracy, 96.59% sensitivity, and 97.9% specificity for two kinds of leukemia are classified.

Furthermore, authors set out to design a deep feedforward algorithm, in order to classify the provided microarray cancer data into a set of classifications for future diagnosis purposes like in (Basavegowda & Dagnew, 2020). For every dataset they have employed a seven-layer deep neural network design with different settings. Eight commonly used microarray cancer datasets are used to validate the suggested strategy, which involves scaling the feature values using the Min-Max method. On four datasets Leukemia, Lung-Michigan, Ovarian, and Prostate the classification accuracy is 1.00, indicating faultless classification performance, With 0.99 accuracy on Lung-Harvard2, 0.96 accuracy on CNS and colon, and 0.95 accuracy on breast cancer.

However, in (Zulfiqar et al., 2022) the author suggested building a strong deep learning model to identify Geobacter pickeringiis 4mC sites. The predicted model encoded the DNA sequences of Geobacter pickeringii using two types of feature descriptors, the binary and k-mer composition. Correlation and an incremental feature selection (IFS) approach combined with a gradient-boosting decision tree (GBDT)-based algorithm were used to improve the

merged features. Subsequently, the refined characteristics were introduced into a 1D convolutional neural network (CNN) with the purpose of distinguishing 4mC sites in Geobacter pickeringii from non-4mC sites. The predicted model performed an accuracy of 0.868.

While, the authors in (Saxena et al., 2022) proposed a novel approach by used Symmetric uncertainty, Principal Component Analysis, Mean imputation, K-nearest neighbour, Random forest, Decision trees, and Neural network methods to obtained of Decision trees accuracy about 76.07, Random forest accuracy (79.8), Multilayer perceptron accuracy (77.60), and K-nearest neighbour accuracy is (78.58).

Also, in (Mahendran & P M, 2022) the authors are suggested a approach to used preprocessing techniques were used to improve classification ability, quality control, normalization, and downstream analysis were performed on DNA methylation data by Random forest, LASSO, SVM embedded feature selection through DNA methylation data was used for the analysis and get the results about quality control eliminated poorly performing samples with p-values of 0.01, data was normalized using log2 transformation and Z-scores, differentially methylated positions (DMPs) were determined using a fold change (FC) of 2 and p-value of 0.01, ada Boost selected CpG sites with the highest accuracy of 87%, and 12 CpG sites were selected by Ada Boost during the 3rd fold.

However, the authors in (Alhenawi et al., 2022) proposed an ensemble FS methods (Homogeneous and Heterogeneous), combination processes (union, intersection, voting), and thresholding processes (static thresholds, complexity measures) through using a Hybrid FSM, Wrapper-based FSM, Filter FS, and Parallel FS methods with attained the results of DFS provides a successful rate that equals 57 and 18 improvement rate compared to traditional methods, the proposed method provides an average accuracy above 90% using SVM, KNN, and C4.5 classifiers, the proposed method provides an average accuracy that equals 92%, and the proposed S model (EU) gives 100% accuracy over 3 out of 5 datasets, it contribute is hybrid approach aims to improve classification accuracy without affecting computation time.

While, the authors in (Kishore et al., 2023) used the methods include SMOTE and SMOTE followed by random undersampling for class imbalance, three pipelines of hybrid feature selection techniques: mRMR followed by CFS, mRMR, mutual information followed by CFS, and mRMR followed by SVM-RFE. Class balancing using SMOTE and random undersampling, CNN model for class balancing using SMOTE, DNN model for overall macro-average AUC score with using TCGA, METABRIC datasets to predict of IDC breast cancer, then obtained the results as accuracy measure and Cohen Kappa score were used for multi-class classification, The AutoKeras generated model exhibited the highest accuracy post SMOTE and post SMOTE and random undersampling.

Finaly, in (Biswas et al., 2023) the authors are suggested to building a potential machine learning model to predict heart disease by using Linear regression (LR), Decision tree (DT), Naive Bayes (NB), Random forest (RF), Support vector machine (SVM), K-nearest neighbour (KNN), and Artificial neural network (ANN), it gets the results from its analysis by using a processed dataset for feature selection and classification tasks, and obtained the results of highest accuracy (94.51) achieved by C4 for SF3, C1 had the second highest accuracy (93.41) for all three SFs, C2 had poor accuracy (75.82) for SF3, C4 had low accuracy (78.02 and 76.92) for SF1 and SF2, other algorithms had accuracy between 84.61 and 92.31, best algorithm for the dataset is C4 for SF3.

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

Table 1. Main Characteristics of Surveyed Studies

Ref.	Year	Dataset	Techniques	Result & Accuracy	Contribution
(Mohammed & Abdulazeez, 2017)	2017	Microarray, Hypoxia	- Mahalanobis distance, Partition around medoids (PAM)	- Optimal cluster solution for Hypoxia dataset when k=3, Dunn's index value of 0.0783 for both Hypoxia and ATMs datasets, Improved cluster quality with proposed algorithm using Weighted and Normalized Mahalanobis distance, Highest Dunn's index value for ATMs dataset when k=2, Enhanced performance of PAM algorithm with proposed distance on ATMs dataset	Identification of relevant gene patterns in microarray data, Proposal of an enhanced PAM algorithm based on weighted Normalised Mahalanobis distance, Improvement of cluster quality in microarray expression data using the proposed algorithm.
(Zeebaree et al., 2018)	2018	Ten Cancer Datasets	Multilayered CNN (Convolutional Neural Network) algorithm, Deep learning algorithm	Proposed CNN had a mean classification accuracy of 94.74, Best accuracy performance of proposed CNN was 100, mSVM-RFE-IRF had a mean classification accuracy of 85.82, Best accuracy	Proposed CNN scored higher accuracies compared to other methods in cancer datasets, k-NN classifier performed better than random forest in accurate classification, Random Survival Forest

Ref.	Year	Dataset	Techniques	Result & Accuracy	Contribution
				performance of mSVM-RFE-IRF was 95.55	strategy for selecting informative genes, PSOC4.5 hybrid used for classifying informative genes with superior accuracy.
(H. Al-Baity & Al-Mutlaq, 2021)	2020	GE, DNA	SVM DT RF	The SVM achieved the lowest execution time of 0.02s and the highest accuracy 99.77% on the GE dataset.	The algorithm can decrease the possibility of mistakes and speed up the examination of medical data.
(Kang et al., 2019b)	2019	MLL, Lymphoma, Brain, TOX_171, CNS, DLBCL, Lung	SVM	The classification accuracy is improved, on the DLBCL, CNS, Lung, Ovarian, Lymphoma, and MLL datasets, it achieves 100%, while on the Brain it 96%, and on TOX_171 is 81.38%	Have a high precision and stops overfitting selecting flexible kernels for nonlinearity strong generalization skills
(Khan et al., 2019)	2019	Leukemia	SVM RF	The accuracy of the proposed method is 0.9980, it is higher than other methods.	The algorithm has a better results and highest accuracy compared to other methods
(Khorshid & Abdulazeez, 2021)	2021	DDSM, MIAS , WDBC	K-NN , ANN , SVM - Logistic Regression (LR), RF, NB classifier, SL	SVM has the highest accuracy of 98% in image processing techniques, LR	Machine learning and mechanistic modeling for prediction of

Ref.	Year	Dataset	Techniques	Result & Accuracy	Contribution
			strategies, SVM and K-NN	and LDA have accuracies of 97.23% and 95.73% respectively, K-NN technique achieved the best results compared to NB and CART, SVM has the highest accuracy of 97.07% among 8 ML algorithms, ANNs have achieved the best accuracy, precision, and F1 score, SVM and RF Classifier are the best predictive analyzes with an accuracy of 96.5%, ANN and CNN have the highest accuracies of 99.3% and 97.3% respectively, SVM cubic classifier has an accuracy of 92.3%, SVM has an accuracy of 96% and K-NN has an accuracy of 100% in breast cancer detection.	metastatic relapse in early-stage breast cancer, Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer, Breast cancer detection using automated whole breast ultrasound, Pectoral muscle segmentation for cancer detection and diagnosis.
(Cho et al., 2022)	2022	Bladder cancer	SVM RF DNN	The result show that DNN models yielded more significant	Enhanced the accuracy by using ML

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

Ref.	Year	Dataset	Techniques	Result & Accuracy	Contribution
				precision and AUCs than SVM models, despite SVM marginally superior test accuracy and MCC.	classification algorithm
(Singh et al., 2020)	2020	ALLAML, CLL SUB, GLI, GLIOMA Prostate-GE, SMK-CAN	DNN	FsNet achieve high accuracy with a significantly smaller number of parameters	Provide a high-dimensional data with small number of samples
(Sun et al., 2019b)	2019	Brain-tumor, Colon, Prostate, Lung, Leukemia, DLBCL, SRBCT, 9-Tumor	KNN C4.5 SVM	FSDNE algorithm produced the highest average classification accuracy of 84% using the KNN, then C4.5 and SVM classifiers less accuracy.	Enhance the neighborhood decision systems capacity for classification and decision-making
(Wu & Hicks, 2021)	2021	Breast cancer	SVM	SVM algorithm have the best performance with an accuracy of 90%, a recall of 87%, and a specificity of 90%	The SVM algorithm provide high specificity, recall, and accuracy in differentiating between breast cancer subtypes point to its potential utility in clinical settings
(Zhang et al., 2021)	2021	Microarray gene expression for many cancer type	SVM	SVM achieved the highest accuracy of 99.58%	a framework for feature selection that can effectively extract features from high-dimensional

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

Ref.	Year	Dataset	Techniques	Result & Accuracy	Contribution
					cancer gene expression datasets
(M. Ghosh et al., 2020b)	2019	Monk1, Monk2 Breast Cancer, Wine Horse, Ionosphere Soybean-small Arrhythmia, Hill-valley, Madelon	WFACOFS KNN MLP SVM	WFACOFS algorithm has the best accuracy for eight of the ten datasets	It is focuses on building a multi-objective FS algorithm based on ACO, to improve the accuracy and feature reduction
(Tabl et al., 2019b)	2019	Breast cancer gene	SVM BNB RF	Bayesian Naive Bayes give the highest accuracy	Extending and developing the idea of guided learning
(Khalifa et al., 2020b)	2020	Cancer	DT CNN DL	The suggested method produced a 96.90% total testing accuracy.	Offer a high degree of accuracy in classification techniques by limiting the number of characteristics to the best ones and eliminating the unnecessary ones.
(Bouke et al., 2023)	2023	Cancer types based on gene expression	BukaGini	The BukaGini algorithm find important feature interactions across a range of datasets and enhancing the model performance.	Improves feature selection based on the Gini index.
(Kurniabudi et al., 2020)	2020	CICIDS-2017	RF BN RT	The J48 algorithm has 99.87% accuracy	To improve the efficiency of anomaly/attack

Ref.	Year	Dataset	Techniques	Result & Accuracy	Contribution
			NB J48	using 52 relevant selected characteristics with a longer time, whereas the RF has accuracy of 99.86% utilizing the relevant selected features of 22.	detection, they identify the most important and pertinent features.
(P. Ghosh et al., 2021b)	2021	Heart disease	AB DT GB KNN RF	The results show that RFBM performs especially well with high impact features and generates accuracy that is significantly greater than previous work, using 10 features with accuracy of 99.05%.	Provide a reliable technique to as precisely anticipate cardiac disease as feasible.
(Shaban et al., 2020b)	2020	COVID-19 Patients	KNN HSFM	The accuracy of the suggested CPDS was 96%, surpassing that of other contemporary techniques.	Compared to current approaches, the suggested CPDS strategy performs better and introduces the best detection accuracy with the least amount of time penalty.
(Xia et al., 2021)	2021	Microarray	RF	Using only 20% of the characteristics in the incredibly vast microarray	The RF method improved the accuracy

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

Ref.	Year	Dataset	Techniques	Result & Accuracy	Contribution
				datasets, the method decreased their dimensionality and confirmed the genetic pathways.	
(Seal et al., 2020b)	2020	DNA methylation CNV	DL	With a 95.1% classification accuracy, the features that the DDAE extracted have demonstrated good classification ability.	It can classify diseases type by using ML algorithm also improved the accuracy
(Mallick et al., 2023)	2020	Microarray data expressions of 72 leukemia patients	DL DNN	The DNN classifier outperforms accuracy about 98%.	The classifying the leukemia data is made simpler and more accurate by the deep learning technique and automated analysis of microarray data
(Basavegowda & Dagnew, 2020)	2019	CNS, Colon, Prostate, Leukaemia, Ovarian, Lung-Harvard2, Lung-Michigan, Breast cancers	DNN	Leukemia, Lung-Michigan, Ovarian, and Prostate the classification accuracy is 1.00, with 0.99 accuracy on Lung-Harvard2, 0.96 accuracy on CNS and colon, and 0.95 accuracy on breast cancer.	Increase the binary datasets classification accuracy

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

Ref.	Year	Dataset	Techniques	Result & Accuracy	Contribution
(Zulfiqar et al., 2022)	2022	DNA	CNN DT	The predicted model performed an accuracy of 0.868	Increase the performance of the predicted model using ML
(Saxena et al., 2022)	2022	Microarray	PCA, K-NN, Random Forest, Decision trees, Neural network	Decision trees accuracy about 76.07, Random forest accuracy (79.8), Multilayer perceptron accuracy (77.60), and K-nearest neighbour accuracy is (78.58).	Feature selection was done using correlation attribute, information gain, and principal component analysis methods. The number of features selected for classification was six. Parameters were optimized for the classification model.
(Mahendran & P M, 2022)	2022	DNA methylation	Random forest, LASSO, SVM, Logistic regression (LR)	Quality control eliminated poorly performing samples with p-values of 0.01, Data was normalized using log2 transformation and Z-scores, Differentially methylated positions (DMPs) were determined using a fold change (FC) of 2	The paper discusses the use of machine learning, deep learning, and advanced statistical and mathematical algorithms. The paper suggests that early identification of AD is crucial for the development of a cure.

Ref.	Year	Dataset	Techniques	Result & Accuracy	Contribution
				and p-value of 0.01, Ada Boost selected CpG sites with the highest accuracy of 87%, 12 CpG sites were selected by Ada Boost during the 3rd fold.	
(Alhenawi et al., 2022)	2022	Microarray	Hybrid FSM, Wrapper-based FSM, Filter FS, Parallel FS methods	DFS provides a successful rate that equals 57 and 18 improvement rate compared to traditional methods, The proposed method provides an average accuracy above 90% using SVM, KNN, and C4.5 classifiers, The proposed method provides an average accuracy that equals 92%, The proposed S model (EU) gives 100% accuracy over 3 out of 5 datasets.	Hybrid approach aims to improve classification accuracy without affecting computation time.
(Kishore et al., 2023)	2023	TCGA, METABRIC	mRMR, CFS, MI, SVM-RFE	Accuracy measure and Cohen Kappa score were used for multi-class classification,	machine learning algorithms have been developed to prognosticate

International Journal of Research and Applied Technology

5(1)(2025) 104-138

Journal homepage: https://ojs.unikom.ac.id/index.php/injuratech

Ref.	Year	Dataset	Techniques	Result & Accuracy	Contribution
				The AutoKeras generated model exhibited the highest accuracy post SMOTE and post SMOTE and random undersampling. The CNN model showed the highest CKS in both class balancing techniques.	the stage and classification of cancer; however, there has been a dearth of endeavors to preprocess the gene expression data, employ deep learning methodologies, and ascertain the stage of cancer with utmost precision.
(Biswas et al., 2023)	2023	Heart disease	LR, DT, NB, RF, SVM, KNN, Artificial neural network (ANN)	Highest accuracy (94.51) achieved by C4 for SF3, C1 had the second highest accuracy (93.41) for all three SFs, C2 had poor accuracy (75.82) for SF3, C4 had low accuracy (78.02 and 76.92) for SF1 and SF2, Other algorithms had accuracy between 84.61 and 92.31, Best algorithm for the dataset is C4 for SF3.	K.A. and M.A.M. provided the idea and designed the experiments, all authors discussed the results and contributed to the manuscript.

6. Discussion

A comparison of the publications that use machine learning for feature selection in gene selection analysis is presents in the above table. The best way for finding highest dataset

accuracy is use machine learning algorithms for feature selection because. All above methods in the table have been depended on machine learning algorithm to get highest accuracy, like SVM, RF, CNN, DT, KNN, DL, BN, all of them tried to provide best performance, and at the least amount of time penalty to classified the dataset for patients that is made simpler and easy way to determine the disease in a short time. In general, the analyzed methods accuracy varies from one approach to the next. However, compared to the conventional techniques, the optimization-based feature selection approach that made use of machine learning algorithms performed better. The methods most used classifier is SVM. Multiple classifiers were employed in certain papers. Nevertheless, in those papers that employed several classifiers, the SVM obtained superior accuracy than the others. Additionally, the outperformance shows that, there are differences in processing times and feature selection techniques play a crucial part in providing high accuracy and optimal performance when detecting gene expression for various diseases.

7. Conclusion

The paper provides a comprehensive review of feature selection methods in gene expression analysis, it categorizes feature selection strategies into filter, wrapper, and embedded methods. - Popular feature selection algorithms such as Relief, RFE, and Mutual Information are discussed. The performance of these methods is evaluated across different datasets and conditions, emerging trends in feature selection, such as the integration of domain knowledge and deep learning approaches, are addressed. The paper emphasizes the importance of developing robust and interpretable feature selection methods, the challenges and future directions in the field are discussed.

Acknowledgement

I extend my heartfelt gratitude to all those who contributed to the successful completion of this comprehensive review on "Feature Selection Methods of Gene Expression Based on Machine Learning". This endeavor would not have been possible without the support, guidance, and encouragement of numerous individuals and resources.

I would like to express my sincere appreciation to my supervisor: Prof. Eng. Dr. Adnan Mohsin Abdulazeez, whose insightful supervision and expertise played a pivotal role in shaping the direction of this review. Your guidance and feedback were invaluable throughout the research process.

Finally, I express my appreciation to all the authors of the studies reviewed and the scientific community at large. Their collective efforts contribute to the advancement of knowledge and inspire researchers like myself to explore new frontiers in the intersection of gene expression and machine learning.

References

Abdulqader, D. M., Abdulazeez, A. M., & Zeebaree, D. Q. (2020). Machine Learning Supervised Algorithms of Gene Selection: A Review. 62(03).

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. Electronics, 9(8), 1295. https://doi.org/10.3390/electronics9081295

Ahmed, O., & Brifcani, A. (2019). Gene Expression Classification Based on Deep Learning. 2019 4th Scientific International Conference Najaf (SICN), 145-149. https://doi.org/10.1109/SICN47020.2019.9019357

Al-Azzam, N., & Shatnawi, I. (2021). Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer. Annals of Medicine and Surgery, 62, 53-64. https://doi.org/10.1016/j.amsu.2020.12.043

Alhenawi, E., Al-Sayyed, R., Hudaib, A., & Mirjalili, S. (2022). Feature selection methods on gene expression microarray data for cancer classification: A systematic review. Computers in Biology and Medicine, 140, 105051. https://doi.org/10.1016/j.compbiomed.2021.105051

Ali, M., & Aittokallio, T. (2019). Machine learning and feature selection for drug response prediction in precision oncology applications. Biophysical Reviews, 11(1), 31-39. https://doi.org/10.1007/s12551-018-0446-z

Almazrua, H., & Alshamlan, H. (2022). A Comprehensive Survey of Recent Hybrid Feature Selection Methods in Cancer Microarray Gene Expression Data. IEEE Access, 10, 71427-71449. https://doi.org/10.1109/ACCESS.2022.3185226

Almugren, N., & Alshamlan, H. (2019). A Survey on Hybrid Feature Selection Methods in Microarray Gene Expression Data for Cancer Classification. IEEE Access, 7, 78533-78548. https://doi.org/10.1109/ACCESS.2019.2922987

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. Decision Analytics Journal, 3, 100071. https://doi.org/10.1016/j.dajour.2022.100071

Basavegowda, H. S., & Dagnew, G. (2020). Deep learning approach for microarray cancer data classification. CAAI Transactions on Intelligence Technology, 5(1), 22-33. https://doi.org/10.1049/trit.2019.0028

Berry, M. W., Mohamed, A., & Yap, B. W. (Eds.). (2020). Supervised and Unsupervised Learning for Data Science. Springer International Publishing. https://doi.org/10.1007/978-3-030-22475-2

Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, Md. R., Azam, S., Ahmed, K., Bui, F. M., Al-Zahrani, F. A., & Moni, M. A. (2023). Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. BioMed Research International, 2023, 1-15. https://doi.org/10.1155/2023/6864343

Bommert, A., Welchowski, T., Schmid, M., & Rahnenführer, J. (2022). Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. Briefings in Bioinformatics, 23(1), bbab354. https://doi.org/10.1093/bib/bbab354

Bouke, M. A., Abdullah, A., Frnda, J., Cengiz, K., & Salah, B. (2023). BukaGini: A Stability-Aware Gini Index Feature Selection Algorithm for Robust Model Performance. IEEE Access, 11, 59386-59396. https://doi.org/10.1109/ACCESS.2023.3284975

Burkart, N., & Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. Journal of Artificial Intelligence Research, 70, 245-317. https://doi.org/10.1613/jair.1.12228

Ceriotti, M. (2019). Unsupervised machine learning in atomistic simulations, between predictions and understanding. The Journal of Chemical Physics, 150(15), 150901. https://doi.org/10.1063/1.5091842

Chen et al. - 2020−A novel selective naïve Bayes algorithm.pdf. (n.d.).

Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. Knowledge-Based Systems, 192, 105361. https://doi.org/10.1016/j.knosys.2019.105361

Cho, H., Tong, F., You, S., Jung, S., Kim, W. H., & Kim, J. (2022). Prediction of the Immune Phenotypes of Bladder Cancer Patients for Precision Oncology. IEEE Open Journal of Engineering in Medicine and Biology, 3, 47–57. https://doi.org/10.1109/OJEMB.2022.3163533

Ferdous, M., Debnath, J., & Chakraborty, N. R. (2020). Machine Learning Algorithms in Healthcare: A Literature Survey. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1–6. https://doi.org/10.1109/ICCCNT49239.2020.9225642

Ghosh, M., Guha, R., Sarkar, R., & Abraham, A. (2020a). A wrapper-filter feature selection technique based on ant colony optimization. Neural Computing and Applications, 32(12), Article 12. https://doi.org/10.1007/s00521-019-04171-3

Ghosh, M., Guha, R., Sarkar, R., & Abraham, A. (2020b). A wrapper-filter feature selection technique based on ant colony optimization. Neural Computing and Applications, 32(12), 7839–7857. https://doi.org/10.1007/s00521-019-04171-3

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. M. J. M., Ignatious, E., Shultana, S., Beeravolu, A. R., & De Boer, F. (2021a). Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques. IEEE Access, 9, 19304–19326. https://doi.org/10.1109/ACCESS.2021.3053759

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. M. J. M., Ignatious, E., Shultana, S., Beeravolu, A. R., & De Boer, F. (2021b). Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques. IEEE Access, 9, 19304–19326. https://doi.org/10.1109/ACCESS.2021.3053759

H. Al-Baity, H., & Al-Mutlaq, N. (2021). A New Optimized Wrapper Gene Selection Method for Breast Cancer Prediction. Computers, Materials & Continua, 67(3), 3089–3106. https://doi.org/10.32604/cmc.2021.015291

Haq, A. U., Li, J. P., Saboor, A., Khan, J., Wali, S., Ahmad, S., Ali, A., Khan, G. A., & Zhou, W. (2021). Detection of Breast Cancer Through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques. IEEE Access, 9, 22090–22105. https://doi.org/10.1109/ACCESS.2021.3055806

Hassan, N. S., Abdulazeez, A. M., Zeebaree, D. Q., & Hasan, D. A. (2021). Medical Images Breast Cancer Segmentation Based on K-Means Clustering Algorithm: A Review. Asian Journal of Research in Computer Science, 23–38. https://doi.org/10.9734/ajrcos/2021/v9i130212

Hossain, Md. A., Saiful Islam, S. M., Quinn, J. M. W., Huq, F., & Moni, M. A. (2019). Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality. Journal of Biomedical Informatics, 100, 103313. https://doi.org/10.1016/j.jbi.2019.103313

Itoo, F., Meenakshi, & Singh, S. (2021). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology, 13(4), 1503–1511. https://doi.org/10.1007/s41870-020-00430-y

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. Behavior Therapy, 51(5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002

Jo, T. (2021). Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning. Springer International Publishing. https://doi.org/10.1007/978-3-030-65900-4

Kang, C., Huo, Y., Xin, L., Tian, B., & Yu, B. (2019a). Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. Journal of Theoretical Biology, 463, 77–91. https://doi.org/10.1016/j.jtbi.2018.12.010

Kang, C., Huo, Y., Xin, L., Tian, B., & Yu, B. (2019b). Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. Journal of Theoretical Biology, 463, 77–91. https://doi.org/10.1016/j.jtbi.2018.12.010

Kegerreis, B., Catalina, M. D., Bachali, P., Geraci, N. S., Labonte, A. C., Zeng, C., Stearrett, N., Crandall, K. A., Lipsky, P. E., & Grammer, A. C. (2019). Machine learning approaches to predict lupus disease activity from gene expression data. Scientific Reports, 9(1), 9617. https://doi.org/10.1038/s41598-019-45989-0

Khalifa, N. E. M., Taha, M. H. N., Ezzat Ali, D., Slowik, A., & Hassanien, A. E. (2020a). Artificial Intelligence Technique for Gene Expression by Tumor RNA-Seq Data: A Novel Optimized Deep Learning Approach. IEEE Access, 8, 22874–22883. https://doi.org/10.1109/ACCESS.2020.2970210

Khalifa, N. E. M., Taha, M. H. N., Ezzat Ali, D., Slowik, A., & Hassanien, A. E. (2020b). Artificial Intelligence Technique for Gene Expression by Tumor RNA-Seq Data: A Novel Optimized Deep Learning Approach. IEEE Access, 8, 22874–22883. https://doi.org/10.1109/ACCESS.2020.2970210

Khan, Z., Naeem, M., Khalil, U., Khan, D. M., Aldahmani, S., & Hamraz, M. (2019). Feature Selection for Binary Classification Within Functional Genomics Experiments via Interquartile Range and Clustering. IEEE Access, 7, 78159–78169. https://doi.org/10.1109/ACCESS.2019.2922432

Khorshid, S. F., & Abdulazeez, A. M. (2021). BREAST CANCER DIAGNOSIS BASED ON K-NEAREST NEIGHBORS: A REVIEW.

Kishore, A., Venkataramana, L., Prasad, D. V. V., Mohan, A., & Jha, B. (2023). Enhancing the prediction of IDC breast cancer staging from gene expression profiles using hybrid feature selection methods and deep learning architecture. Medical & Biological Engineering & Computing, 61(11), 2895–2919. https://doi.org/10.1007/s11517-023-02892-1

Kurniabudi, Stiawan, D., Darmawijoyo, Bin Idris, M. Y., Bamhdi, A. M., & Budiarto, R. (2020). CICIDS-2017 Dataset Feature Analysis With Information Gain for Anomaly Detection. IEEE Access, 8, 132911–132921. https://doi.org/10.1109/ACCESS.2020.3009843

Lindholm, A., Wahlström, N., Lindsten, F., & Schön, T. B. (n.d.). Supervised Machine Learning.

Liu, H., Zhou, M., & Liu, Q. (2019). An embedded feature selection method for imbalanced data classification. IEEE/CAA Journal of Automatica Sinica, 6(3), 703–715. https://doi.org/10.1109/JAS.2019.1911447

Liu, X., Zhu, X., Li, M., Wang, L., Zhu, E., Liu, T., Kloft, M., Shen, D., Yin, J., & Gao, W. (2018). Multiple Kernel k-means with Incomplete Kernels.

Mahendran, N., & P M, D. R. V. (2022). A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease. Computers in Biology and Medicine, 141, 105056. https://doi.org/10.1016/j.compbiomed.2021.105056

Mahesh, B. (2018). Machine Learning Algorithms−A Review. 9(1).

Mallick, P. K., Mohapatra, S. K., Chae, G.-S., & Mohanty, M. N. (2023). Convergent learning-based model for leukemia classification from gene expression. Personal and Ubiquitous Computing, 27(3), 1103–1110. https://doi.org/10.1007/s00779-020-01467-3

Maniruzzaman, Md., Jahanur Rahman, Md., Ahammed, B., Abedin, Md. M., Suri, H. S., Biswas, M., El-Baz, A., Bangeas, P., Tsoulfas, G., & Suri, J. S. (2019). Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. Computer Methods and Programs in Biomedicine, 176, 173–193. https://doi.org/10.1016/j.cmpb.2019.04.008

Mohammed, N. N., & Abdulazeez, A. M. (2017). Gene clustering with partition around mediods algorithm based on weighted and normalized mahalanobis distance. 2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), 140–145. https://doi.org/10.1109/ICIIBMS.2017.8279707

Ngiam, K. Y., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. The Lancet Oncology, 20(5), e262–e273. https://doi.org/10.1016/S1470-2045(19)30149-4

Pinal-Fernandez, I., Casal-Dominguez, M., Derfoul, A., Pak, K., Miller, F. W., Milisenda, J. C., Grau-Junyent, J. M., Selva-O'Callaghan, A., Carrion-Ribas, C., Paik, J. J., Albayda, J., Christopher-Stine, L., Lloyd, T. E., Corse, A. M., & Mammen, A. L. (2020). Machine learning algorithms reveal unique gene expression profiles in muscle biopsies from patients with different types of myositis. Annals of the Rheumatic Diseases, 79(9), 1234–1242. https://doi.org/10.1136/annrheumdis-2019-216599

Rahat, A. M., Kahir, A., & Masum, A. K. M. (2019). Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset. 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), 266–270. https://doi.org/10.1109/SMART46866.2019.9117512

Ray, S. (n.d.). A Quick Review of Machine Learning Algorithms.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Saxena, R., Sharma, S. K., Gupta, M., & Sampada, G. C. (2022). A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods. Computational Intelligence and Neuroscience, 2022, 1–11. https://doi.org/10.1155/2022/3820360

Scheurer, M. S., & Slager, R.-J. (2020). Unsupervised Machine Learning and Band Topology. Physical Review Letters, 124(22), 226401. https://doi.org/10.1103/PhysRevLett.124.226401

Seal, D. B., Das, V., Goswami, S., & De, R. K. (2020a). Estimating gene expression from DNA methylation and copy number variation: A deep learning regression model for multi-omics integration. Genomics, 112(4), Article 4. https://doi.org/10.1016/j.ygeno.2020.03.021

Seal, D. B., Das, V., Goswami, S., & De, R. K. (2020b). Estimating gene expression from DNA methylation and copy number variation: A deep learning regression model for multi-omics integration. Genomics, 112(4), 2833–2841. https://doi.org/10.1016/j.ygeno.2020.03.021

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. In J. K. Mandal & D. Bhattacharya (Eds.), Emerging Technology in Modelling and Graphics (Vol. 937, pp. 99–111). Springer Singapore. https://doi.org/10.1007/978-981-13-7403-6_11

Shaban, W. M., Rabie, A. H., Saleh, A. I., & Abo-Elsoud, M. A. (2020a). A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. Knowledge-Based Systems, 205, 106270. https://doi.org/10.1016/j.knosys.2020.106270

Shaban, W. M., Rabie, A. H., Saleh, A. I., & Abo-Elsoud, M. A. (2020b). A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. Knowledge-Based Systems, 205, 106270. https://doi.org/10.1016/j.knosys.2020.106270

Shokrzade, A., Ramezani, M., Akhlaghian Tab, F., & Abdulla Mohammad, M. (2021). A novel extreme learning machine based kNN classification method for dealing with big data. Expert Systems with Applications, 183, 115293. https://doi.org/10.1016/j.eswa.2021.115293

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. IEEE Access, 8, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Singh, D., Climente-González, H., Petrovich, M., Kawakami, E., & Yamada, M. (2020). FsNet: Feature Selection Network on High-dimensional Biological Data (arXiv:2001.08322). arXiv. http://arxiv.org/abs/2001.08322

Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. Fco. (2020). A review of unsupervised feature selection methods. Artificial Intelligence Review, 53(2), 907–948. https://doi.org/10.1007/s10462-019-09682-y

Srinivasa, K. G., Siddesh, G. M., & Manisekhar, S. R. (Eds.). (2020). Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications. Springer Singapore. https://doi.org/10.1007/978-981-15-2445-5

Sun, L., Zhang, X., Qian, Y., Xu, J., & Zhang, S. (2019a). Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. Information Sciences, 502, 18–41. https://doi.org/10.1016/j.ins.2019.05.072

Sun, L., Zhang, X., Qian, Y., Xu, J., & Zhang, S. (2019b). Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. Information Sciences, 502, 18–41. https://doi.org/10.1016/j.ins.2019.05.072

Surya and Subbulakshmi–2019–Sentimental Analysis using Naive Bayes Classifier.pdf. (n.d.).

Tabl, A. A., Alkhateeb, A., ElMaraghy, W., Rueda, L., & Ngom, A. (2019a). A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. Frontiers in Genetics, 10. https://www.frontiersin.org/articles/10.3389/fgene.2019.00256

Tabl, A. A., Alkhateeb, A., ElMaraghy, W., Rueda, L., & Ngom, A. (2019b). A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. Frontiers in Genetics, 10. https://www.frontiersin.org/articles/10.3389/fgene.2019.00256

Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019). A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 1255–1260. https://doi.org/10.1109/ICCS45141.2019.9065747

Tharwat, A. (2021). Independent component analysis: An introduction. Applied Computing and Informatics, 17(2), 222–249. https://doi.org/10.1016/j.aci.2018.08.006

Toğaçar, M., Ergen, B., Cömert, Z., & Özyurt, F. (2020). A Deep Feature Learning Model for Pneumonia Detection Applying a Combination of mRMR Feature Selection and Machine Learning Models. IRBM, 41(4), 212–222. https://doi.org/10.1016/j.irbm.2019.10.006

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19(1), 281. https://doi.org/10.1186/s12911-019-1004-8

Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. A., Elkhatib, Y., Hussain, A., & Al-Fuqaha, A. (2019). Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges. IEEE Access, 7, 65579–65615. https://doi.org/10.1109/ACCESS.2019.2916648

Wu, J., & Hicks, C. (2021). Breast Cancer Type Classification Using Machine Learning. Journal of Personalized Medicine, 11(2), 61. https://doi.org/10.3390/jpm11020061

Xia, H., Akay, Y. M., & Akay, M. (2021). Selecting Relevant Genes From Microarray Datasets Using a Random Forest Model. IEEE Access, 9, 97813–97821. https://doi.org/10.1109/ACCESS.2021.3092368

Xing, W., & Bei, Y. (2020). Medical Health Big Data Classification Based on KNN Classification Algorithm. IEEE Access, 8, 28808–28819. https://doi.org/10.1109/ACCESS.2019.2955754

Yuan, A., You, M., He, D., & Li, X. (2022). Convex Non-Negative Matrix Factorization With Adaptive Graph for Unsupervised Feature Selection. IEEE Transactions on Cybernetics, 52(6), 5522–5534. https://doi.org/10.1109/TCYB.2020.3034462

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. J, 2(2), 226–235. https://doi.org/10.3390/j2020016

Yuan, F., Lu, L., & Zou, Q. (2020). Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 1866(8), 165822. https://doi.org/10.1016/j.bbadis.2020.165822

Zeebaree, D. Q., Haron, H., & Abdulazeez, A. M. (2018). Gene Selection and Classification of Microarray Data Using Convolutional Neural Network. 2018 International Conference on Advanced Science and Engineering (ICOASE), 145–150. https://doi.org/10.1109/ICOASE.2018.8548836

Zhang, J., Xu, D., Hao, K., Zhang, Y., Chen, W., Liu, J., Gao, R., Wu, C., & De Marinis, Y. (2021). FS-GBDT: Identification multicancer-risk module via a feature selection algorithm by integrating Fisher score and GBDT. Briefings in Bioinformatics, 22(3), bbaa189. https://doi.org/10.1093/bib/bbaa189

Zulfiqar, H., Huang, Q.-L., Lv, H., Sun, Z.-J., Dao, F.-Y., & Lin, H. (2022). Deep-4mCGP: A Deep Learning Approach to Predict 4mC Sites in Geobacter pickeringii by Using Correlation-Based Feature Selection Technique. International Journal of Molecular Sciences, 23(3), 1251. https://doi.org/10.3390/ijms23031251