Literature Review

1. Relief-F

- **Algorithm/Tools**: Relief-F, F-score, SVM
- **Dataset**: Wisconsin Diagnostic Breast Cancer Dataset
- **Results**: The Relief-F method is commonly used for feature selection in classification problems. By estimating the quality of features based on how well their values distinguish between instances that are near each other but belong to different classes, it achieved a high accuracy of 96.7%. It also demonstrated improved stability and feature relevance, making it a robust method for handling noisy datasets.
- **Reference**: Haury et al., 2011(

SpringerLink

)

2. Chi-Square Filter

- **Algorithm/Tools**: Chi-square, ANOVA
- **Dataset**: Text classification datasets
- **Results**: Chi-square and ANOVA are filter-based methods used to evaluate the correlation between features and the target variable. These methods showed improved precision, especially when dealing with imbalanced data, making them suitable for text classification tasks.
- **Reference**: Bahassine et al., 2020(

SpringerLink

)

3. MRMR (Minimum Redundancy Maximum Relevance)

- **Algorithm/Tools**: Mutual Information, Support Vector Machine (SVM)
- **Dataset**: Lung cancer classification using CT images
- **Results**: MRMR focuses on selecting features that are highly relevant to the target variable while minimizing redundancy among the features themselves. This method reduced feature redundancy and enhanced interpretability in lung cancer classification, contributing to better diagnostic accuracy.
- **Reference**: Togaçar et al., 2020(

SpringerLink

)

4. LASSO (Least Absolute Shrinkage and Selection Operator)

- **Algorithm/Tools**: Lasso regression
- **Dataset**: Genomic datasets (multi-class classification)
- **Results**: LASSO is a popular method for feature selection and regularization, especially in high-dimensional datasets. By applying Lasso regression, it selected the most relevant genes in genomic datasets, increasing classification accuracy significantly. This is crucial for biological applications where selecting a subset of important features (genes) is essential.
- **Reference**: Ang et al., 2016(

SpringerLink

)

5. PCA (Principal Component Analysis)

- **Algorithm/Tools**: PCA, K-NN (K-Nearest Neighbors)
- **Dataset**: UCI Machine Learning Repository (various datasets)
- **Results**: PCA is a widely used dimensionality reduction technique that projects data into a lower-dimensional space while retaining as much variance as possible. In this case, it reduced the dimensionality of various datasets by 30% while maintaining classification accuracy. This is particularly beneficial for improving computational efficiency.
- **Reference**: Kaur et al., 2019(

SpringerLink

)

6. Boruta

- **Algorithm/Tools**: Random forest, Feature importance
- **Dataset**: Cardiovascular disease dataset
- **Results**: Boruta is a feature selection algorithm based on random forests. It helps identify all relevant features by comparing the importance of original features to that of shadow features. It identified 95% of relevant features, significantly improving the precision and recall of cardiovascular disease predictions.
- **Reference**: Kou et al., 2020(

SpringerLink

)

7. Wrapper Methods

- **Algorithm/Tools**: Genetic Algorithms, Naive Bayes
- **Dataset**: Breast cancer, spam detection
- **Results**: Wrapper methods use a search algorithm to evaluate subsets of features based on a machine learning model's performance (in this case, Naive Bayes). These methods achieved

increased overall precision and recall, as well as a reduction in false positives, making them effective for datasets like breast cancer diagnosis and spam detection.

- **Reference**: Mesut et al., 2020(

SpringerLink

)

8. Hybrid Methods

- **Algorithm/Tools**: Relief-F + Wrapper, ANOVA + SVM
- **Dataset**: HIV-1 Protease Cleavage Site Prediction
- **Results**: Hybrid methods combine filter and wrapper approaches to leverage the strengths of both. This combination enhanced prediction capabilities in the HIV-1 dataset while reducing computational time compared to using only wrapper methods.
- **Reference**: Ötürk et al., 2013(

SpringerLink

)

9. Ensemble Methods

- **Algorithm/Tools**: Particle Swarm Optimization (PSO), Neural Networks
- **Dataset**: Gene expression data, protein datasets
- **Results**: Ensemble methods combine multiple feature selection algorithms to improve robustness and accuracy. In this case, using PSO and neural networks for feature selection resulted in a high classification accuracy of 98% for gene expression datasets.
- **Reference**: Huda et al., 2021(

SpringerLink

)

10. Filter-Based Methods

- **Algorithm/Tools**: t-Test, Chi-square, ANOVA
- **Dataset**: Financial fraud detection datasets
- **Results**: Filter methods independently assess each feature's relationship with the target variable. For financial fraud detection, these methods showed an improvement in fraud detection accuracy by 4%, making them suitable for scenarios with a high volume of irrelevant data.
- **Reference**: Ravisankar et al., 2011(

SpringerLink

)

11. Consistency-Based Methods

- **Algorithm/Tools**: Consistency-based subset selection
- **Dataset**: Microarray gene expression datasets
- **Results**: Consistency-based methods aim to find the smallest subset of features that provides the same classification consistency as the original dataset. In this case, the method achieved 85% consistency with reduced feature dimensionality.
- **Reference**: Dash & Liu, 2003(

SpringerLink

)

12. Correlation-Based Methods

- **Algorithm/Tools**: Feature Interaction, Correlation Coefficients
- **Dataset**: Text categorization, Sentiment Analysis
- **Results**: Correlation-based methods evaluate the redundancy and relevance of features based on their correlation with the target variable. These methods achieved higher feature relevance and improved the performance of text classification tasks.
- **Reference**: Tang et al., 2019(

SpringerLink

)

13. Mutual Information-Based Methods

- **Algorithm/Tools**: Granular multi-label feature selection
- **Dataset**: Multi-label image datasets (bioimaging, pattern recognition)
- **Results**: Mutual information measures the statistical dependency between variables. Using this method, the study achieved 94% accuracy in multi-label classification, making it highly effective for tasks like bioimaging and pattern recognition.
- **Reference**: Li et al., 2017(

SpringerLink

)

14. Deep Learning-Based Methods

- **Algorithm/Tools**: Neural networks, Auto-encoders
- **Dataset**: Medical imaging datasets (MRI, X-ray, CT)
- **Results**: Deep learning-based methods such as neural networks and autoencoders are increasingly used for feature selection in complex datasets. In medical imaging, this approach achieved a 99% accuracy for cancer detection, combining feature selection with powerful classifiers like Convolutional Neural Networks (CNNs).
- **Reference**: Liu et al., 2018(

SpringerLink

)

15. Feature Dependence Methods

- **Algorithm/Tools**: Dependence maximization (HSIC-based)
- **Dataset**: UCI datasets (cancer, diabetes, heart disease)
- **Results**: Feature dependence methods like HSIC-based techniques maximize the statistical dependence between features and the target variable. They enhanced classification accuracy by over 8% compared to traditional feature selection methods, demonstrating their effectiveness in various medical and health-related datasets.
- **Reference**: Song et al., 2012(

SpringerLink

)

S.No.	Method	Algorithm/Tools	Dataset	Results	Reference
1	Relief-F	Relief-F, F-score, SVM	Wisconsin Diagnostic Breast Cancer Dataset	Achieved 96.7% accuracy, higher stability and feature relevance	Haury et al., 2011(SpringerLink)
2	Chi-Square Filter	Chi-square, ANOVA	Text classification datasets	Improved precision for imbalanced data	Bahassine et al., 2020(SpringerLink)
3	MRMR	Mutual Information, Support Vector Machine	Lung cancer classification (CT images)	Reduced redundancy and improved interpretability	Togaçar et al., 2020(SpringerLink)
4	LASSO	Lasso regression	Genomic datasets (multi-class classification)	Increased classification accuracy by selecting most relevant genes	Ang et al., 2016(SpringerLink)
5	PCA	Principal Component Analysis (PCA), K-NN	UCI Machine Learning Repository (various datasets)	Reduced dimensionality by 30%, maintaining similar classification accuracy	Kaur et al., 2019(SpringerLink)
6	Boruta	Random forest, Feature importance	Cardiovascular disease dataset	Identified 95% relevant features, improving precision and recall	Kou et al., 2020(SpringerLink)
7	Wrapper Methods	Genetic Algorithms, Naive Bayes	Breast cancer, spam detection	Increased overall precision and recall, reduced false positives	Mesut et al., 2020(SpringerLink)
8	Hybrid Methods	Relief-F + Wrapper, ANOVA + SVM	HIV-1 Protease Cleavage Site Prediction	Enhanced prediction capabilities with reduced computational time	Ötürk et al., 2013(SpringerLink

S.No.	Method	Algorithm/Tools	Dataset	Results	Reference
					)
9	Ensemble Methods	PSO (Particle Swarm Optimization), Neural Networks	Gene expression data, protein datasets	Achieved 98% classification accuracy with ensemble feature selection	Huda et al., 2021(SpringerLink)
10	Filter-Based Methods	t-Test, Chi-square, ANOVA	Financial fraud detection datasets	Improved fraud detection accuracy (by 4%)	Ravisankar et al., 2011(SpringerLink)
11	Consistency-Based	Consistency-based subset selection	Microarray gene expression datasets	Achieved 85% consistency with reduced feature dimensionality	Dash & Liu, 2003(SpringerLink)
12	Correlation-Based	Feature Interaction, Correlation Coefficients	Text categorization, Sentiment Analysis	Achieved higher feature relevance and selection for text classification	Tang et al., 2019(SpringerLink)
13	Mutual Information-Based	Granular multi-label feature selection	Multi-label image datasets (bioimaging, pattern recognition)	Achieved 94% accuracy, effective for multi-label classification	Li et al., 2017 (SpringerLink)
14	Deep Learning-Based	Neural networks, Auto-encoders	Medical imaging datasets (MRI, X-ray, CT)	99% accuracy for cancer detection using CNNs and autoencoders combined with feature selection	Liu et al., 2018(SpringerLink)
15	Feature Dependence	Dependence maximization (HSIC-based)	UCI datasets (cancer, diabetes, heart disease)	Enhanced classification accuracy by over 8% compared to traditional methods	Song et al., 2012(SpringerLink)