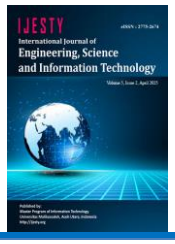
International Journal of Engineering, Science and Information Technology
Volume 5 No. 2 (2025) pp. 33-41
ISSN 2775-2674 (online)
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v5i2.777
Research Paper, Short Communication, Review, Technical Paper



Cataract Eye Disease Diagnosis Using the Random Forest Method

Lilis Novita*, Wahyu Fuadi, Kurniawati

Department of Informatics, Faculty of Engineering, Universitas Malikussaleh, Aceh, Indonesia

**Corresponding author Email: lilis.200170238@mhs.unimal.ac.id*

The manuscript was received on 10 June 2024, revised on 1 September 2024, and accepted on 27 January 2025, date of publication 1 April 2025

Abstract

This study developed a machine learning-based classification model using the Random Forest algorithm to detect cataract risk based on 10 variables, including age, family history, lens opacity, visual acuity decline, light sensitivity, colour changes, double vision, intraocular pressure, slit-lamp examination results, and visual acuity. The model achieved an accuracy of 95.0%, precision of 100%, sensitivity of 92.86%, F1 Score of 96.30%, and specificity of 100% after optimization using Grid Search. Feature importance analysis revealed that lens opacity and visual acuity were the most significant factors in predicting cataract risk, followed by intraocular pressure and visual acuity decline. The system was implemented using Google Colab for model training and Streamlit for an interactive interface, enabling real-time predictions with intuitive visualization of results. This system is expected to be an efficient and high-accuracy supporting tool for medical professionals in early cataract diagnosis.

Keywords: *Cataract, Classification, Data Mining, Machine Learning, Random Forest.*

1. Introduction

As one of the primary human senses, the eyes play a crucial role in performing daily activities. The primary function of the eyes is as a light-detecting mechanism, enabling humans to distinguish between light and dark while forming visual perceptions, which allow us to see our surroundings. The retina is one of the most critical components of the eye because it converts light impulses into nerve signals, which are then interpreted by the brain as the images we see [1]. This process underpins human visual ability. However, eye diseases can disrupt visual function and make everyday activities difficult. Eye health problems, such as cataracts, glaucoma, and diabetic retinopathy, can range from mild to severe. Cataracts, for instance, occur when the eye's lens becomes cloudy, obstructing the passage of light and resulting in blurred vision. Various factors, including oxidation, exposure to ultraviolet radiation, heredity, and nutrition, can cause this condition. Diagnosis is the process of identifying weaknesses or medical conditions experienced by an individual through careful examination and analysis of the symptoms presented [2].

Data mining is the process of manually analyzing data to obtain additional information that was previously undetected, which will then be stored in a specialized database. This process involves extracting significant patterns from the available data to create models or functions that describe and differentiate the concepts or labels within the data [3]. Data mining techniques are widely used across various industries. The application of data mining in the healthcare sector has become a new and popular model because, through these techniques, we can extract information and uncover hidden patterns in data that can serve as a basis for decision-making. One commonly used method in data mining is classification, which aims to predict related variables. Classification algorithms such as Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine are examples of algorithms used in this process. By utilizing these classification algorithms, we can make initial predictions about a patient's condition, which can aid in efforts to prevent premature death [4].

Classification, a form of data analysis, aids in predicting class labels for given samples. Various classification techniques have emerged in machine learning, expert systems, and statistics. Since the 1990s, researchers have curated software repositories to enhance their understanding of data. The evaluation of decision tree methods (C4.5), k-nearest neighbour (k-NN), Naive Bayes, Random Forest, and Decision Stump, by comparing algorithms like C4.5, Random Forest, and SimpleCART, revealed that C4.5 performed best for systems

[5]. Furthermore, considering the decline in classification accuracy for categorical and numerical attributes, out of 20 methods, Bayes Net, Naive Bayes, Classification via Regression, Logistic Regression, and Random Forest were the top choices. Naive Bayes, Bayes Net, and Random Forest proved the most effective for mixed-attribute datasets. Regression classification, NBTree, and multiclass classification were the preferred methods for numerical attribute datasets. Similarly, NBTree, Classification via Regression, and Bayes Net demonstrated superior performance for categorical attribute datasets. While methods such as PART and Decision Tree emerged as the best based on the outlined rules, it is crucial to note that the effectiveness of classification techniques varies depending on the nature of the data. No single classifier universally outperforms others across all datasets [6].

The Random Forest technique is built upon the Decision Tree approach by utilizing multiple decision trees, with each tree trained on different samples and splitting attributes randomly selected from a subset. This technique offers several advantages, such as improved accuracy when handling missing data, effectively managing outliers, and being efficient in data storage. Additionally, this method incorporates a feature selection mechanism to identify optimal features, thereby enhancing the performance of the classification model.

This study applies the Random Forest method for diagnosing cataract eye disease. This method has proven effective in classifying medical data and predicting health conditions by combining multiple decision trees trained on different samples. Random Forest enhances diagnostic accuracy by effectively handling missing data and outliers and identifying optimal features to improve classification model performance. Previous studies have demonstrated the success of this technique in various healthcare contexts, including the early detection of disease symptoms.

2. Literature Review

2.1. Eye

The eye is a sensory organ found in animals and humans that functions to receive light and convert it into electrochemical signals sent to the brain via the optic nerve. The eye serves as a sensory tool in humans. It constantly adjusts the amount of light entering, focuses on objects near and far, and produces continuous images immediately transmitted to the brain [7]. The eye has its functions, which are interconnected with each other. The functions of the eye include:

1. Cornea: The cornea is a transparent tissue that forms the outer layer and covers the front part of the eye. It functions as a lens that allows light to enter the eye.
2. Aqueous Humor: Aqueous humour is a clear fluid that helps deliver nutrients to the eye's tissues.
3. Iris: The iris is responsible for regulating the amount of light entering the eye by adjusting the size of the pupil.
4. Pupil: The pupil is the black circular opening in the centre of the eye. It functions as an opening through which light can enter the eye.
5. Sclera: The sclera is the white, tough tissue surrounding the entire eye. It is the attachment point for the eye muscles that allow the eye to move.
6. Lens: The lens is the second part of the eye after the cornea that helps focus light and images onto the retina. The lens comprises transparent, flexible tissue behind the iris and pupil.
7. Vitreous: The vitreous has a gel-like structure that fills the eye's back part. It plays a role in maintaining the shape of the eye and keeping the retina in place.
8. Retina: The retina is a layer of light-sensitive tissues located on the inner surface of the eye. It processes incoming light into electrical signals sent to the brain through the optic nerve.
9. Choroid: The choroid is a dark brown membrane between the sclera and the retina. It contains blood vessels that supply blood and nutrients to the retina and other eye parts.
10. Eye Muscles: These muscles are responsible for controlling the lens size and supporting the crystalline lens.

2.2. Eye Diseases

The eye is one of the five senses that play a vital role in human life, specifically as the organ of vision [8]. When disorders or eye diseases occur, the impact is very concerning, and if not treated seriously, it can significantly affect a person's quality of life. Therefore, maintaining eye health is very important in your daily routine. Below are several types of eye diseases:

1. Miopia: Myopia, or nearsightedness, is when a person has difficulty seeing distant objects clearly, while close-up vision is relatively good. This condition occurs because the focal point of light entering the eye falls too close to the lens, causing the light to focus incorrectly on the retina. Myopia can be corrected with glasses, contact lenses, or vision correction surgery such as LASIK.
2. Hypermetropia: Hypermetropia, or farsightedness, is when a person has difficulty seeing close-up objects clearly, while distant vision may still be fine. This condition occurs when the eye is too short, or the lens is too flat, causing the light entering the eye to not focus on the retina. Hypermetropia can be corrected with glasses or contact lenses that help to refocus the light correctly.
3. Astigmatism: Astigmatism occurs when the eye's cornea is not perfectly round or the lens has an irregular shape. As a result, light entering the eye cannot focus properly on the retina, causing blurred or distorted vision. Astigmatism can be corrected with glasses, contact lenses, or vision correction surgery.
4. Presbyopia: Presbyopia is a condition that typically occurs in older age. It happens due to the loss of elasticity of the eye's lens due to ageing, which affects the ability to see close-up objects. Presbyopia can be corrected with reading glasses or multifocal contact lenses.
5. Cataract: Cataract is a condition where the lens becomes cloudy, blocking light from entering the retina. This can cause blurred or distorted vision. Cataracts are typically related to ageing but can also be caused by genetic factors, eye injuries, or medical conditions.
6. Glaucoma: Glaucoma is an eye disorder that damages the optic nerve. It is usually associated with increased pressure in the eye, which can lead to gradual vision loss without apparent early symptoms. Glaucoma can be treated with medication, laser procedures, or surgery to reduce eye pressure.
7. Conjunctivitis: Conjunctivitis is the inflammation of the conjunctiva or the lining behind the eyelids and eyeballs. This eye condition occurs when the eyes become red, itchy, watery, and burning. It often affects both children and adults. Several factors that cause this condition include viral or bacterial infections, allergies (such as pollen, dust, wind, or smoke), prolonged use of contact

lenses, and poor hygiene. Treatment usually involves eye drops containing anti-allergy or antibiotic medications, depending on the underlying cause.

8. Blepharitis: Blepharitis is inflammation of the eyelid that causes the eyes to feel itchy red, and scales are on the eyelids. Blepharitis can be caused by bacterial infections or dysfunction of the oil glands in the eyelids. Treatment includes proper cleaning, warm compresses, and prescribed eye ointments or eye drops.
9. Keratitis: Keratitis is inflammation or irritation in the eye's cornea. Eye injuries or infections are the primary causes of keratitis. It can occur due to infectious causes, such as bacterial (mainly Streptococcus, Pseudomonas, Enterobacter), viral (such as Herpes Simplex and Herpes Zoster), or parasitic infections. Non-infectious causes include dry eyes, medication toxicity, contact lens irritation, and allergies. If left untreated or not adequately managed, keratitis can worsen and lead to complications.
10. Pterygium: Pterygium is an eye condition caused by the growth of tissue that forms a triangular, reddish mass on the white part of the eyeball. This condition is known as 'surfer's eye' or 'pinguecula.' It usually starts on the cornea near the nose and can extend to the pupil (the black part of the eye). Typically, it affects only one eye, but if it spreads to both eyes, it is called pterygium. This condition can cause discomfort and itching. If not treated promptly, it may increase the risk of corneal scarring and vision problems. Treatment options include steroid eye drops, lubricating eye drops, and other related medications.

2.3. Data Mining

Data mining is an analytical step in discovering knowledge within a database, known as Knowledge Discovery in Database (KDD). The knowledge can include previously unknown data patterns or relationships between data. The KDD process consists of several stages: data collection, transformation, and analysis [9]. Data mining can generally be categorized into descriptive and predictive [10]. Descriptive data mining is utilized to uncover patterns that are understandable to humans and explain the characteristics of the data.

On the other hand, predictive data mining is employed to create a knowledge model that can be used to make predictions. Based on its functionalities, data mining is divided into seven categories: clustering, which involves grouping data based on similarities; classification, which assigns data to predefined categories; forecasting, which predicts future trends or behaviours; regression, which models the relationship between variables; association, which identifies relationships between data items; sequencing, which analyzes sequences or patterns over time; and descriptive, which summarizes and interprets the characteristics of data. By combining disciplines such as statistics, artificial intelligence, and machine learning, data mining has become a valuable tool for analyzing data and extracting useful information from large datasets [11].

2.4. Classification

Classification is the process of grouping objects based on specific characteristics. It is a training process (learning) of a function with a target used to separate each group of objects into a predefined class label. Classification falls under the type of analysis that can help determine the class label for several samples that need to be classified [12]. This classification technique is suitable for describing datasets with data types such as binary or nominal. However, the limitation of this technique is that it is not ideal for ordinal data sets due to the implicit order in the categories of data [13].

2.5. Random Forest

Random Forest is an extension of the decision tree method that uses multiple decision trees, where each decision tree is trained using individual samples, and each attribute is selected randomly from the set of available attributes. It has several advantages, such as improving accuracy when there is missing data and handling outliers effectively and efficiently storing data. It also features a feature selection process, where it can choose the best features, which in turn can enhance the performance of the classification model [14].

Random Forest uses bagging with random feature selection for each built model. This ensemble method in machine learning is developed to improve the stability and accuracy of the algorithms used in classification and prediction. Bagging can also reduce variance and help prevent overfitting [15].

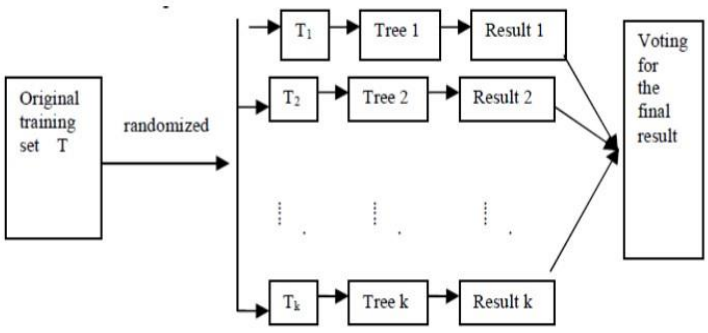


Fig 1. Random Forest Mechanism

For each node in the decision tree of a random forest, the process begins by randomly selecting a subset from the set of attributes at the node and then choosing the optimal attribute to partition the data. This is done using the formula:

$$k = log_2 d \quad (1)$$

Where:

k : Controlling the level of randomness introduced.

d : Number of Attribute.

2.5. C4.5 Algorithm

The C4.5 algorithm is used for classification or clustering in a dataset [16]. Developed by Ross Quinlan, C4.5 is an enhancement of the ID3 algorithm and was designed to improve classification performance more efficiently and accurately. This algorithm is highly popular in data analysis and is used to predict the class or category of data based on existing attributes. It is also one of the most commonly used decision tree techniques and generates a decision tree that is easier to understand [17].

The steps for performing calculations with the C4.5 algorithm include preparing the training data, determining the tree's root by calculating entropy, then calculating the gain, and selecting the tuple to partition. To calculate entropy, the formula is used as follows [18]:

$$Entropy(s) = -\sum_{i-1}^{n} pi \times \log(pi) \quad (2)$$

Where:
s : Set of cases
n : Number of Samples
pi : Proportion of the class

After calculating the entropy, the next step is to calculate the gain using the following equation:

$$Gain(s,A) = Entropy(s) - \sum_{i=1}^{n} \frac{|si|}{|s|} \times \log(pi) \quad (3)$$

Where:
s : Set of Cases
A : Attribute
n : Number of partitions of attribute A
|si| : Proportion of Si relative to S
|s| : Set of Cases

After obtaining the gain and entropy values to choose the attribute as the root, select the attribute with the highest gain ratio from the available attributes. Then, calculate the gain ratio using the following equation:

$$gain\ ratio(s,A) = \frac{Gain(s,A)}{Splitinfo(s,A)} \quad (4)$$

Where:
s : Set of Cases
A : Attribute
$Gain(s,A)$: Gain Information of Attribute A
$Splitinfo(s,A)$: Split Information of Attribute A

$$Split\ info(s,A) = \sum_{i-1}^{n} \left(\frac{|si|}{|s|} log_2 \frac{|si|}{|s|}\right) \quad (5)$$

A : Attribute
n : Number of Partitions of Attribue
|s| : Data Number of Attribute
|si| : Data of Partition of Attribute I

2.6. Python

Python is a high-level programming language developed by Guido van Rossum and first released in 1991. Python features an easy-to-write syntax, and it also has a comprehensive library and strong community support because it is open source. This language is well-known for its simple and easy-to-understand syntax, making it very suitable for beginners [19].

Python is easier to understand due to its simple code writing; it is available for free and open source, flexible because it can run on almost all operating systems, and versatile because it can be implemented in web development, mobile apps, and desktop apps. Python is one of the high-level programming languages that is interpreted, interactive, and object-oriented. This language can operate on almost all platforms, including UNIX-based systems, Mac, Windows, and others [20].

3. Research Methods

The research flow is a series of systematic steps the researcher follows in conducting a study. This flow helps ensure that the research is carried out in a structured, logical, and valid manner.

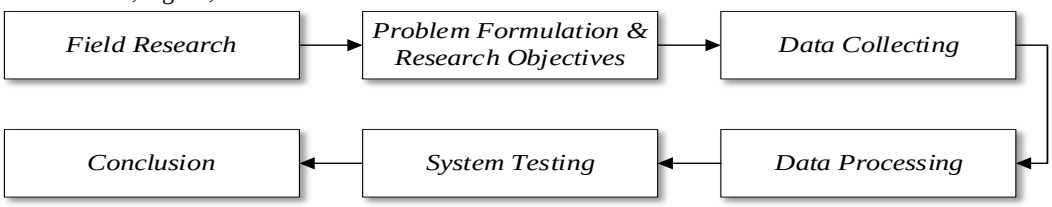


Fig 2. Research Flow Diagram

This diagram illustrates the systematic flow of a research process starting from Field Research, where direct data collection or observations are conducted. The insights from this stage lead to Problem Formulation & Research Objectives, which clearly define the goals and scope of the study. Subsequently, data collecting involves gathering relevant data, refined and analyzed during the data processing phase. The processed data undergoes System Testing, where the system's performance and accuracy are evaluated to validate the findings. Finally, the research concludes with the Conclusion, summarizing the outcomes and insights from the entire process. This structured flow ensures that each step contributes logically and effectively to the research objectives.

3.1. Data Collecting

The dataset used in this study was obtained from patient medical records at the Aceh Tamiang Regional General Hospital from 2019 to 2023. This dataset includes several key variables relevant to the cataract diagnosis process, such as age, medical history, visual acuity (sharpness of vision), light sensitivity, vision decline, vision changes, and intraocular pressure (eye pressure). Additionally, slit-lamp results and visual acuity were recorded for each patient as the primary indicators for detecting cataract conditions. Each row of data contains information on whether a patient was diagnosed with cataracts, labelled as "YES" or "NO" for classification purposes.

3.2. Dataset Overview

Table 1. Initiate Variable

Variable Name	Variabel Code
Age	X1
Riwayat Keluarga	X2
Kekeruhan Lensa	X3
Penurunan Ketajaman	X4
Sensivitas Cahaya	X5
Perubahan Warna	X6
Penglihatan Ganda	X7
Tekanan Intraokular	X8
Hasil Slitlamp	X9
Visus	X10

The "Variable Initialization" table lists variable names used in the data analysis related to eye health, specifically in cataract studies. Each variable is assigned a unique code for easier identification, ranging from demographic factors such as Age (X1) to clinical factors like Lens Opacification (X3), Intraocular Pressure (X8), and Slit-Lamp Results (X9). Measurement variables such as Visual Acuity (X10) are also included. This systematic approach of assigning variable codes facilitates interpretation and calculations during analysis. The dataset used in this study, based on this variable initialization, is as follows:

Table 2. Dataset Overview

NO	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
1	69	0	3.6	8	0	1	0	18	1	0.13
2	63	0	5	6	6	1	0	17.5	0	0.25
3	71	0	0.8	4	0	1	0	21.1	1	0.25
4	80	0	8.5	9	8	0	1	16.2	0	0.27
5	62	0	4.4	6	0	1	0	20.7	0	0.23
…	…	…	…	…	…	…	…	…	…	…
146	72	0	6.9	7	10	0	0	16.5	0	0.27
147	52	0	3.1	6	5	1	1	15.6	0	0.15
148	51	0	5.2	7	9	0	0	13.4	0	0.26
149	70	0	8	2	1	1	0	18.3	0	0.19
150	67	1	0.1	5	4	0	0	14.9	0	0.04

This data will be processed using the Random Forest (RF) algorithm with a total of 150 records to analyze patterns within the dataset and improve accuracy in cataract detection. Several medical variables, such as family history, changes in vision colour, double vision, and slit-lamp examination results, will be converted from categorical to numerical values to enable processing by the algorithm. By leveraging information from these variables, the RF algorithm is expected to provide reliable and consistent diagnostic results, supporting efforts to improve healthcare service quality and the effectiveness of cataract disease management.

4. Result and Discussion

4.1. Import Libary

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns
import joblib
```

Fig 3. Import Library

This code imports various libraries for data analysis, modelling, and evaluating machine learning model performance. Pandas and NumPy are used for data manipulation and computation, while the train_test_split and GridSearchCV modules from sklearn.model_selection assists in splitting the dataset and performing model parameter optimization. RandomForestClassifier is used as the primary model for classification, and LabelEncoder helps convert categorical data into numerical values. For model evaluation, metrics such as accuracy_score, precision_score, recall_score, f1_score, and confusion_matrix are used to measure model performance. matplotlib.pyplot and seaborn are used for data visualization and result plotting, and joblib is used to save the model for future use.

4.2. Model Preparation

```
# Encode variabel kategorikal
le = LabelEncoder()
categorical_columns = ['jenis_kelamin', 'riwayat_keluarga', 'perubahan_warna',
                       'penglihatan_ganda', 'hasil_slitlamp', 'katarak']
for col in categorical_columns:
    df[col] = le.fit_transform(df[col])

# Memisahkan fitur dan label
X = df.drop(columns='katarak')
y = df['katarak']

# Membagi data latih (80%) dan data uji (20%)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Parameter grid untuk C4.5-like Random Forest
param_grid = {
    'n_estimators': [100, 200, 300],  # Jumlah pohon
    'max_depth': [None, 10, 20, 30],  # Kedalaman maksimum pohon
    'min_samples_split': [2, 5, 10],  # Minimum sampel untuk split
    'min_samples_leaf': [1, 2, 4],    # Minimum sampel di leaf
    'criterion': ['entropy'],          # Menggunakan entropy (C4.5) bukan gini
    'max_features': ['sqrt', 'log2'],  # Jumlah fitur yang dipertimbangkan
    'class_weight': ['balanced'],      # Menyeimbangkan kelas
    'bootstrap': [True],               # Menggunakan bootstrap sampling
}

# Inisialisasi model Random Forest
rf_model = RandomForestClassifier(random_state=42)
```

Fig 4. Model Preparation

This code is divided into several sections for data preparation and model initialization. First, categorical variables (riwayat_keluarga, perubahan_warna, penglihatan_ganda, hasil_slitlamp, and katarak) are encoded into numerical values using LabelEncoder, which is necessary for the machine learning model to process the data. Then, the dataset is split into feature variables (X) and the target variable (y), where X contains all the features except for katara, and y contains only the katara column. The data is then divided into training (80%) and test (20%) sets using train_test_split. Next, a param_grid is defined to perform parameter search using Grid Search on the Random Forest model, which resembles C4.5. Parameters such as n_estimators, max_depth, and criterion (with the value entropy) are set. Finally, the Random Forest model is initialized with default parameters using random_state=42 to ensure consistent results.

4.3. Hyperparameter Tuning

```
# Melakukan Grid Search dengan cross-validation
grid_search = GridSearchCV(
    estimator=rf_model,
    param_grid=param_grid,
    cv=5,
    scoring='accuracy',
    n_jobs=-1,
    verbose=2
)

# Melatih model dengan Grid Search
grid_search.fit(X_train, y_train)

# Mendapatkan parameter terbaik
best_params = grid_search.best_params_
print("\nParameter Terbaik:", best_params)

# Menggunakan model terbaik
best_model = grid_search.best_estimator_

# Memprediksi data uji
y_pred = best_model.predict(X_test)

# Menghitung metrik evaluasi
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

# Menghitung Sensitivity dan Specificity
conf_matrix = confusion_matrix(y_test, y_pred)
tn, fp, fn, tp = conf_matrix.ravel()
sensitivity = tp / (tp + fn)
specificity = tn / (tn + fp)
```

Fig 5. Hyperparameter Tuning

This code performs parameter search and model evaluation through several steps. First, GridSearchCV searches for the best parameters with 5-fold cross-validation based on the previously defined parameter grid, where scoring='accuracy' is used as the primary evaluation metric. After the parameter search, the model is trained using the best parameter combination obtained. Then, predictions are made on the test data (y_pred), and evaluation metrics such as accuracy, precision, recall, and F1 score are calculated to measure the model's performance. Additionally, sensitivity and specificity are calculated using the confusion matrix, where sensitivity (tp / (tp + fn)) indicates the model's ability to detect cataracts, and specificity (tn / (tn + fp)) shows the model's ability to identify non-cataract cases correctly.

```
# Menampilkan hasil evaluasi
print("\nHasil Evaluasi Model:")
print(f"Accuracy: {accuracy:.4f}")
print(f"Precision: {precision:.4f}")
print(f"Recall (Sensitivity): {recall:.4f}")
print(f"F1 Score: {f1:.4f}")
print(f"Specificity: {specificity:.4f}")

# Visualisasi Feature Importance
feature_importance = pd.DataFrame({
    'feature': X.columns,
    'importance': best_model.feature_importances_
})
feature_importance = feature_importance.sort_values('importance', ascending=False)

plt.figure(figsize=(10, 6))
sns.barplot(x='importance', y='feature', data=feature_importance)
plt.title('Feature Importance dalam Prediksi Katarak')
plt.xlabel('Importance Score')
plt.tight_layout()
```

Fig 6. Result Visualization

This code displays the model evaluation results and visualizes the feature importance of each feature in the model. First, it prints the model's evaluation metrics, including Accuracy, Precision, Recall (Sensitivity), F1 Score, and Specificity. These metrics provide insight into the model's performance in classifying cataract cases. Next, the code visualizes the importance of the feature from the model, showing the contribution of each feature in predicting cataracts. The feature importance data is stored in a DataFrame and sorted from the most important to the least important features. A bar chart is then generated using Seaborn. Barplot to show the importance scores of each feature, providing an understanding of which features have the most influence on the cataract prediction model.

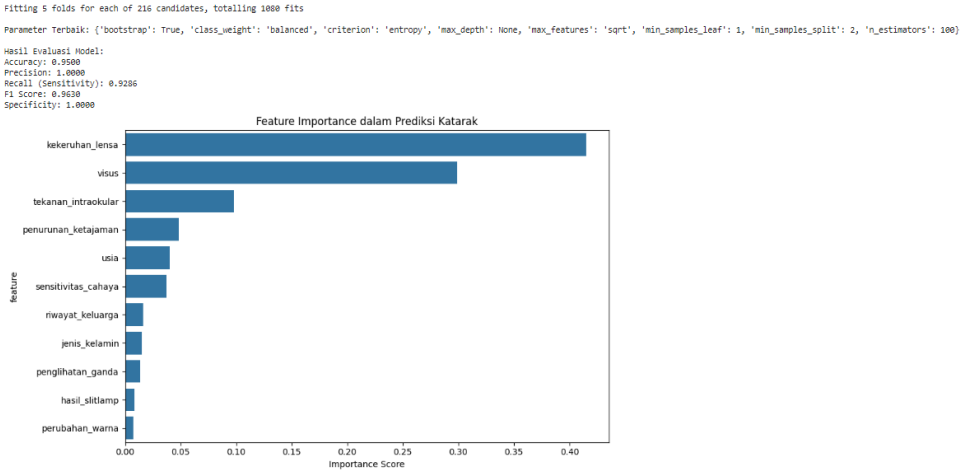


Fig 7. Result in Random Forest Classification

Figure 7 shows the model evaluation results and the feature importance graph in cataract prediction using the Random Forest model. Based on the evaluation results, the model achieved an accuracy of 95%, precision of 100%, recall (sensitivity) of 92.86%, F1 score of 96.30%, and specificity of 100%. This indicates that the model effectively classifies patients with and without cataracts.

In the feature importance graph, lens_opacity and visual_acuity are the most influential features in cataract prediction, followed by intraocular_pressure, visual_acuity_loss, and age. These features contribute more to the model's decisions. Meanwhile, features like double-vision and color_change show relatively more minor influences. This information helps understand the key factors associated with cataract conditions and provides a basis for further future analysis or development of more efficient prediction models.

```
[ ]  # Visualisasi Confusion Matrix
     plt.figure(figsize=(8, 6))
     sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues')
     plt.title('Confusion Matrix')
     plt.xlabel('Predicted')
     plt.ylabel('Actual')
     plt.tight_layout()
```

Fig 8. Confusion Matrix Processed

This code displays a confusion matrix visualization to evaluate the model's prediction results. The confusion matrix is visualized using a seaborn heatmap, where the numbers in each cell represent the number of correct and incorrect predictions for each class (cataract and non-cataract). The graph size is set to 8x6, and each cell is annotated with numbers to show the absolute values of these predictions. The X-axis is labelled "Predicted," and the Y-axis is labelled "Actual," providing context about the model's predictions versus the actual conditions. With the blue colour scheme (cmap='Blues'), this visualization makes it easier to understand the model's performance in correctly classifying patients and identifying prediction errors, such as false positives and false negatives.

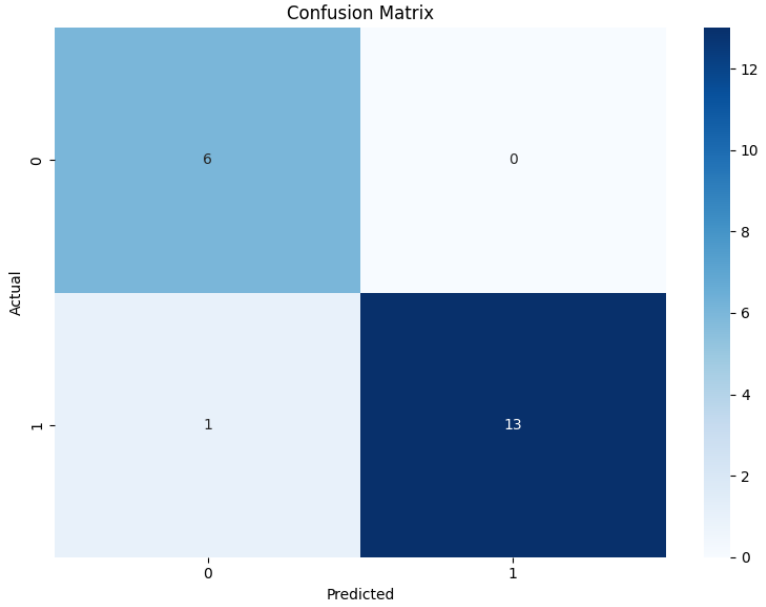


Fig 9. Confusion Matrix Result

Figure 9 displays the confusion matrix of the model's prediction results, where the Y-axis represents the actual values, and the X-axis shows the model's predictions. The matrix consists of four central cells: 6 true negatives (first row, first column), which represent non-cataract patients correctly predicted as non-cataract; 13 true positives (second row, second column), indicating cataract patients correctly predicted as cataract. One false negative (second row, first column), where a cataract patient was incorrectly predicted as non-cataract, and zero false positives (first row, second column), meaning no non-cataract patient was incorrectly predicted as cataract. This matrix indicates that the model has a high accuracy rate with minimal errors, particularly in classifying cataract conditions.

5. Conclusion

Several key conclusions can be drawn based on the research and implementation of the cataract classification system using the Random Forest algorithm. The model achieved an accuracy of 95.0%, demonstrating its reliability in predicting cataract risk based on symptoms and patient examination results. It also achieved a perfect precision of 100%, with no misclassifications in the positive cataract category, and had a sensitivity of 92.86%, F1 Score of 96.30%, and specificity of 100%. The best parameters from Grid Search included bootstrap = True, class_weight = balanced, criterion = entropy, max_depth = None, max_features = sqrt, min_samples_leaf = 1, min_samples_split = 2, and n_estimators = 100. Feature importance analysis revealed that lens opacification and visual acuity were the most significant factors in determining cataract risk, followed by intraocular pressure and visual decline. The system was successfully implemented using Streamlit, enabling users to input patient data interactively and receive fast, accurate cataract predictions. It also featured visualizations that displayed data distribution and diagnostic probability, allowing users or healthcare professionals to interpret the results and understand the influence easily.

References

[1] S. Nurhaningsih, Y. Susanti, and S. Handajani, “Implementasi Algoritma C5.0 Untuk Klasifikas Penyakit Gagal Ginjal Kronik,” 2019.

[2] L. Budhy Adzy and A. Pambudi, “ALGORITMA NAÏVE BAYES UNTUK KLASIFIKASI KELAYAKAN PENERIMA BANTUAN IURAN JAMINAN KESEHATAN PEMERINTAH DAERAH KABUPATEN SUKABUMI,” 2023.

[3] B. W. K. Nurdin, “IMPLEMENTASI DATA MINING UNTUK MENGKLASIFIKASI DATA NASABAH PT. ADIRA FINANCE ACEH TENGAH MENGGUNAKAN ALGORITMA C4.5,” *Jurnal Sistem Informasi Kaputama (JSIK)*, vol. 1, no. 1, 2017.
[4] M. Suhendri and Y. Afrilia, “Klasifikasi Karya Ilmiah (Tugas Akhir) Mahasiswa Menggunakan Metode Naive Bayes Classifier (NBC),” 2021. [Online]. Available: http://sistemasi.ftik.unisi.ac.id
[5] R. Widiyanti, C. Suhery, and R. Hidayati, “Implementasi Algoritma C5.0 Untuk Klasifikasi Kepuasan Masyarakat Terhadap Pelayanan Kantor Kecamatan,” *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 4, p. 1200, Aug. 2022, doi: 10.30865/jurikom.v9i4.4632.
[6] A. P. Ruise, A. S. Mashuri, M. Sulaiman, and F. Rahman, "Studi Komparasi Metode Svm, Logistic Regresion Dan Random Forest Clasifier Untuk Mengklasifikasi Fake News di Twitter," *J I M P - Jurnal Informatika Merdeka Pasuruan*, vol. 7, no. 2, p. 64, Sep. 2023, doi: 10.51213/jimp.v7i2.472.
[7] A. Prabowo, “Sistem pakar untuk mendiagnosa penyakit mata pada manusia menggunakan metode certainty factor,” vol. 3, no. 4, 2023.
[8] D. Meidelfia, F. Sukmaa, and S. Y. Kharismaa, “Analisis Klasifikasi Penyakit Mata dengan Perbandingan Metode Random Forest dan Metode K-Nearest Neighbor,” *Jurnal Internasional Komputasi dan Teknik Tingkat Lanjut*, vol. 5, no. 2, pp. 136–145, 2023.
[9] A. Rofifah, R. Goenjantoro, and D. Yuniarti, “Perbandingan Pengelompokan K-Means dan K-Medoids Pada Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Studi Kasus : Data Titik Panas Di Indonesia Pada 28 April 2018),” *Jurnal EKSPONENSIAL*, vol. 10, no. 2, pp. 143–152, 2019.
[10] A. Yani, Z. Azmi, D. Suherdi, S. Informasi, and S. Triguna Dharma, “Implementasi Data Mining Menganalisa Data Penjualan Menggunakan Algoritma K-Means Clustering,” *JURNAL SISTEM INFORMASI TGD*, vol. 2, no. 2, pp. 315–323, 2023, [Online]. Available: https://ojs.trigunadharma.ac.id/index.php/jsi
[11] C. Zai, “IMPLEMENTASI DATA MINING SEBAGAI PENGOLAHAN DATA,” 2022.
[12] A. Srirahayu and L. Setya Pribadie, “JURNAL ILMIAH INFORMATIKA GLOBAL Review Paper Data Mining Klasifikasi Data Mining,” *Jurnal Ilmiah Informatika Global*, vol. 14, no. 1, 2023.
[13] Fauziah, Dedy Hartama, and Irfan Sudahri Damanik, “Analisa Kepuasan Pelanggan Menggunakan Klasifikasi Data Mining,” *Jurnal Penerapan Kecerdasan Buatan*, 2020.
[14] A. Arisusanto, N. Suarna, and G. Dwilestari, “Analisa Klasifikasi Data Harga Handphone Menggunakan Algoritma Random Forest Dengan Optimize Parameter Grid,” *Jurnal Teknologi Ilmu Komputer*, vol. 1, no. 2, pp. 43–47, 2023, doi: 10.56854/jtik.v1i2.51.
[15] Moch. Lutfi, "Implementasi Metode K-Nearest Neighbor dan Bagging Untuk Klasifikasi Mutu Produksi Jagung," *Agromix*, vol. 10, no. 2, pp. 130–137, 2019, doi: 10.35891/agx.v10i2.1636.
[16] D. A. C, “Implementasi Data Mining Untuk Prediksi Penyakit Diabetes,” vol. 2, no. 1, pp. 39–46.
[17] M. Adriansa, L. Yulianti, L. Elfianty, U. Dehasen Bengkulu, and J. Meranti Raya, “Analisis Kepuasan Pelanggan Menggunakan Algoritma C4.5,” 2022.
[18] P. B. N. Setio, D. R. S. Saputro, and B. Winarno, “Klasifikasi dengan Pohon Keputusan Berbasis Algoritme C4.5,” *Prosiding Seminar Nasional Matematika*, vol. 3, pp. 64–71, 2020, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/
[19] M. R. S. Alfarizi, M. Z. Al-farish, M. Taufiqurrahman, G. Ardiansah, and M. Elgar, “Penggunaan Python Sebagai Bahasa Pemrograman untuk Machine Learning dan Deep Learning,” *Karya Ilmiah Mahasiswa Bertauhid (KARIMAH TAUHID)*, vol. 2, no. 1, pp. 1–6, 2023.
[20] M. Abdul muthalib, I. Irfan, K. Kartika, and S. M. Selamat Meliala, "Pengiraan Pose Model Manusia Pada Repetisi Kebugaran Ai Pemograman Python Berbasis Komputerisasi," *INFOTECH journal*, vol. 9, no. 1, pp. 11–19, 2023, doi: 10.31949/infotech.v9i1.4233.