DNA Profiling: An Investigation of Six Machine Learning Algorithms for Estimating the Number of Contributors in DNA Mixtures

Hamdah Alotaibi[1], Fawaz Alsolami[2]
Department of Computer Science
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah 21589, Saudi Arabia

Rashid Mehmood[3]
High Performance Computing Center
King Abdulaziz University
Jeddah 21589, Saudi Arabia

***Abstract*—DNA (Deoxyribonucleic acid) profiling involves analysis of sequences of individual or mixed DNA profiles to identify persons these profiles belong to. DNA profiling is used in important applications such as for paternity tests, in forensic science for person identification on a crime scheme, etc. Finding the number of contributors in a DNA mixture is a major task in DNA profiling with challenges caused due to allele dropout, stutter, blobs, and noise. The existing methods for finding the number of unknowns in a DNA mixture suffer from issues including computational complexity and accuracy of estimating the number of unknowns. Machine learning has received attention recently in this area but with limited success. Many more efforts are needed for improving the robustness and accuracy of these methods. Our research aims to advance the state-of-the-art in this area. Specifically, in this paper, we investigate the performance of six machine learning algorithms -- Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Stochastic Gradient Descent (SGD), and Gaussian Naïve-Bayes (GNB) -- applied to a publicly available dataset called PROVEDIt, containing mixtures with up to five contributors. We evaluate the algorithmic performance using confusion matrices and four performance metrics namely accuracy, F1-Score, Recall, and Precision. The results show that LR provides the highest Accuracy of 95% for mixtures with five contributors.**

***Keywords*—Machine learning; DNA profiling; DNA mixtures; forensic science**

I. INTRODUCTION

Between different individuals, most of the genome is the same. However, there are some differences, and here comes the science of Deoxyribonucleic acid (DNA) profiling. It is the process that takes benefit from these differences and gives the ability to distinguish between individuals [1]. DNA profiling analyzes DNA sequences that are referred to as genetic markers. The most commonly used genetic marker is Short Tandem Repeats (STRs) [1]. DNA profiling is used in important applications such as for paternity tests, in forensic science for person identification on a crime scheme, etc. [2]. Determining the number of contributors is one of the essential stages in DNA profiling. This task is often not straightforward because of the challenges that could appear, caused due to allele dropout, stutter, blobs, and noise [3], [4].

The current methods for finding the number of unknowns in DNA mixtures can be divided into three types [5]. The first type includes the basic methods which are compute-intensive, are slow, and have accuracy issues (e.g., [6]). The second type includes high-performance computing (HPC) methods, which are faster but highly compute-intensive, and their accuracy requires significant improvements (e.g., [7]). The third type includes machine learning methods that are faster but their classification accuracies and robustness need to be improved, requiring many more efforts in this direction (e.g., [8]).

Recent years have seen rapid and considerable growth in using machine learning in different fields, showing promising results [9]. However, when dealing with inferring the number of contributors in the DNA profile mixture, few researchers have addressed the effect of using machine learning to solve this challenge. To the best of our knowledge, there are three works to date [8], [10], [11], and each one deals with the problem from a different perspective. The research on machine learning based DNA profiling is in its infancy, many more works are needed to improve the diversity and accuracy of the machine learning methods. Our research aims to advance the state-of-the-art in the DNA profiling domain. Specifically, in this paper, we investigate the performance of six machine learning algorithms -- Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Stochastic Gradient Descent (SGD), and Gaussian Naïve-Bayes (GNB) -- applied to a publicly available dataset called PROVEDIt. The dataset contains DNA mixtures with up to five contributors.

We have investigated the performance of these algorithms in detail using four performance metrics namely accuracy, F1-Score, Recall, and Precision. The performance of each algorithm has been analyzed using confusion matrices and graphs of the four matrices for each of the five classes, One-Person, Two-Person, Three-Person, Four-Person, and Five-Person.

For KNN, the highest values for the F1-Score, Recall, and Precision metrics were achieved, all for the Five-Persons class, at 68%, 62%, 75%, respectively. For the RF algorithm, the highest values for the F1-Score, Recall, and Precision metrics were achieved for the Five-Persons class at 86%, One-Person class at 88%, and the Five-Persons class at 90%, respectively.

For SVM, the highest values for the F1-Score, Recall, and Precision metrics were achieved, all for the Five-Persons class, at 96%, 96%, 95%, respectively. For SGD, the highest values for the F1-Score, Recall, and Precision metrics were achieved for the Five-Persons class at 93%, both One-Person and Five-Person classes at 100%, and the Five-Persons class at 88%, respectively. For LR, the highest values for the F1-Score, Recall, and Precision metrics were achieved for the One-Person class at 97%, Five-Persons class at 98%, and the One-Person class at 97%, respectively. For GNB, the highest values for the F1-Score, Recall, and Precision metrics were achieved for the Three-Persons class at 71%, Three-Persons class at 83%, and the Five-Persons class at 100%, respectively. The highest Accuracy over all the algorithms was achieved by the LR algorithm at 95% for mixtures with up to five contributors.

The rest of the paper is organized as follows. Section II briefly reviews the related works. Section III describes the methodology of the proposed work. Section IV presents results and their analyses for the six machine learning algorithms. Section V contains the conclusion and future work.

II. Related Work

The methods for estimating the number of contributors in a DNA mixture can be divided into three types. These are basic methods, HPC methods, and machine learning-based methods. The basic methods and tools include, among others, Maximum Allele Count (MAC) [6], Total Allele Count (TAC) [11], MLE [12], DNA Mixtures [13], Lab Retriever [14] and DNA MIX [15]. The parallel or HPC methods include Euroformix [16], LikeLTD [17] and NOCIt [4], [5], [18]. To the best of our knowledge, only three works have used machine learning to determine the number of contributors in a DNA profile. Since machine learning is the focus of our research, these three methods are reviewed below in some detail.

Marciano and Adelmen [8] evaluated five machine algorithms, and finally, they chose the SVM that reached 98% accuracy in the training stage and 97% accuracy in the testing stage for four contributors. Note that the 97% accuracy is on a dataset with up to four contributors compared to five contributors where typically the accuracy will be lower due to a larger number of classes. The data that they have used consists of 1405 profiles from 20 individuals. Benschop et al. [11] examined ten machine learning algorithms, and finally, they chose the RFC model with 19 features. They used 590 profiles that range from a single person to five person mixtures. They removed both Amelgenin and Y-chromosomal markers. There were more than 250 features for each profile, but they chose only the best 50 features. In terms of Accuracy, they got (83%). Kruijver et al. [10] use decision trees in their work. They used 766 profiles from Global filer multiplex with a 25-second injection. In terms of Accuracy, they got from (77.9% - 85.2%).

The research on machine learning based DNA profiling is in its infancy, many more works are needed to improve the diversity and accuracy of the machine learning methods. Our research aims to advance the state-of-the-art in the DNA profiling domain. Specifically, in this paper, we investigate the performance of six machine learning algorithms.

III. Methodology and Design

This section presents the proposed methodology for this work, depicted in Fig. 1. Section A will give a short explanation of the dataset that has been used. Section B will explain the ML models used in this work, and finally, Section C will show the evaluation metrics used.

A. The Dataset

The data in terms of DNA profiles have been taken from the public dataset PROVEDIt [19]. This dataset contains more than 25,000 STR profiles containing DNA mixtures that range from one to five contributors. The dataset contains more than one kit with different cycles number and injection times. Fig. 2 shows the number of profiles that we have taken from this dataset. We took 156 profiles to represent each class among the five classes, and we ended with 780 DNA profile mixtures, which means that we have 18720 samples (780 profiles * 24 markers). When we collected the data, we made sure it contained different injection times and cycle numbers.

We encountered more than one challenge for the preprocessing stage, including dealing with empty cells, OL values and deleting the unwanted markers. All of these challenges were addressed during the pre-processing phase in order to prepare the dataset for the classification stage.

B. Machine Learning Methods

In this paper, we examined six different machine learning algorithms that are introduced below.

K-Nearest Neighbors (KNN) is considered one of the simplest algorithms in classifying tasks. This algorithm aims to find the samples that exist close to each other [8].

Random Forest (RF) is an algorithm that is used in both classification and regression. As the name implies, it is a set of multiple decision trees. The dataset will be divided into a batch of random datasets, then building a decision tree for each of them. Each decision tree will give a diffident decision, and the majority result will be taken [20].

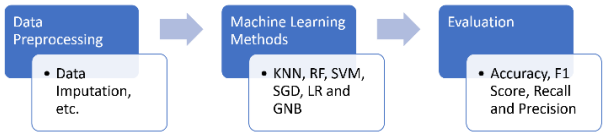

Fig. 1. A High-Level Depiction of our Methodology.

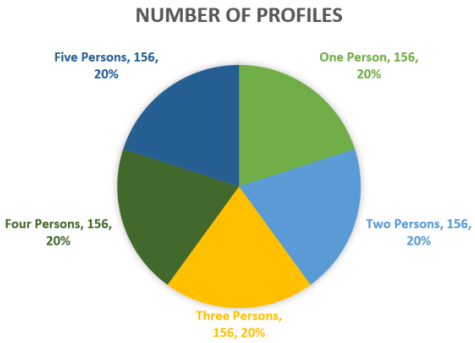

Fig. 2. PROVEDIt: Number and Distribution of DNA Mixtures with the Five Classes (Selected Profiles).

Support Vector Machine (SVM) is a very familiar algorithm when dealing with classification problems. When there is more than one way of drawing the line (boundary) to separate the data points (support vectors), one of the solutions is to measure the distance (margin) between the boundary and the data points. SVM will try to maximize this margin [8].

Stochastic Gradient Descent (SGD) is a suitable choice when having a significant dataset in terms of size and when there is not much computation. For forward pass, it uses a single sample at random and then changes weights [21].

Logistic Regression (LR) calculates the dependent variable based on the independent variable by calculating the errors between the actual data point and the predicted data point by the linear equation, then square the errors, sum them up, and minimize them [8].

Gaussian NB (GNB) comes from the Gaussian distributions that represent the dataset. It is suitable when the dimensionality of the inputs is complex and high. It used the Bayes theorem. It assumes that each feature is independent of other features [22].

C. Evaluation

In this work, we used four different performance metrics. Which are Accuracy that calculated as following $accuracy = (TP + TN)/(TP + FN + TN + FP)$, F1-Score that calculated as following $f1\ score = 2 * (recall * precision)/(recall + precision)$, Recall that calculated as following $recall = TP/(TP + FN)$, and Precision that calculated as following $precision = (TP/(TP + FP)$. Where TP is True Positive, TN is True Negative, FN is False Negative, and FP is False Positive.

IV. Results and Analysis

This section presents the performance for the six algorithms. The six algorithms: KNN, RF, SVM, SGD, LR and GNB are analyzed respectively in Section IV.A to Section IV.F. Section IV.G will show a comparison between all the six algorithms. Section IV.H provides a brief descriptive comparison of our work in this paper with the earlier related works.

A. Nearest Neighbors (KNN)

Fig. 3 shows the confusion matrix for KNN model. There are five classes. The values vary from the minimum (zero) with purple color to the maximum (627) with dark yellow. The matrix could be read as follows. For Two number of unknowns, for instance, there are (502) correct predictions, (181) samples were misclassified as One-Person, (258) samples were misclassified as Three-Persons mixtures, (70) samples were misclassified as the Four-Persons mixtures and (16) samples were misclassified as the Five-Persons mixtures. The results show that One-Persons class have the highest number of correct predictions (627), then Five Persons class with (626), Three Persons class, Four Persons class and finally Two Person class. In terms of mischaracterization, Two Persons class has the highest number of misclassification (525), then Four Persons class (514), then Three Persons class (471), then One Person class (399) and finally Five Persons class (388).

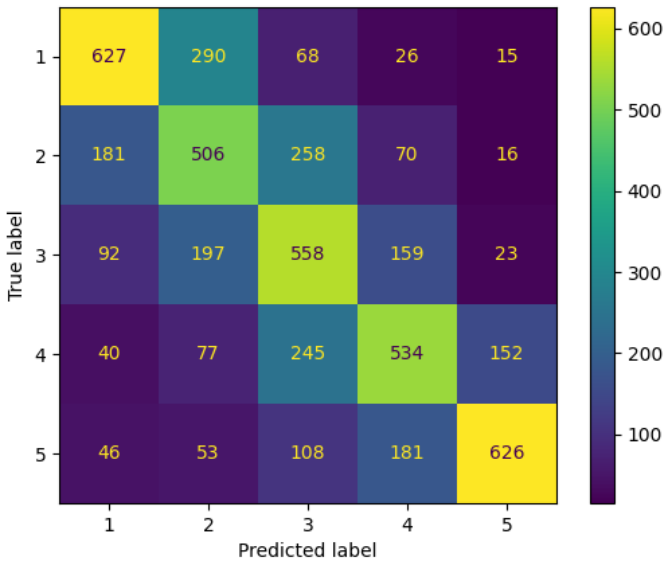

Fig. 3. The Confusion Matrix (KNN).

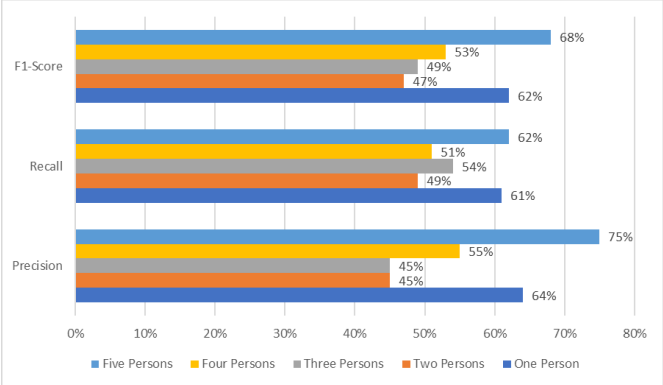

Fig. 4. F1-Score, Recall and Precision (KNN).

Fig. 4 shows F1-Score, Recall and Precision for the KNN model. The highest score is for Five Persons class Precision (75%) because referring to Fig. 3, we know that TP for Five Persons class is 626 and FP is 206, and the lowest is for both Three Persons and Two Persons classes Precision (45%) because as we know TP for Three Persons is (558) and FP is (679), and for Two Persons class TP is (506), and FP is (617). For F1-Score, the highest score is for Five Persons class (68%), and the lowest is for Two Persons class (47%). For Recall, the highest score is for Five Persons class (62%), and the lowest is for Two Persons class (49%). For Precision, the highest score is for Five Persons class (75%), and the lowest is for both Three Persons and Two Persons classes (45%).

B. Random Forest (RF)

Fig. 5 shows the confusion matrix for RF model. There are five classes. The values vary from the minimum (zero) with purple color to the maximum (902) with dark yellow. The matrix could be read as follows. For Three number of unknowns, for instance, there are (808) correct predictions, (41) samples were misclassified as One unknown, (138) samples were misclassified as Two unknown mixtures, (25) samples were misclassified as Four unknown mixtures and (17)

samples were misclassified as Five unknown mixtures. The results show that the One Person class have the highest number of correct predictions (902), then both Five Persons and Four Persons classes with (840), then Two Persons class with (822) and finally Three Persons class with (808). In terms of mischaracterization, Three Persons class has the highest number of misclassification (221), then Two Persons class (209), then Four Persons class (208), then Five Persons class (174) and finally One Person class (124).

Fig. 6 shows F1-Score, Recall and Precision for RF model. The highest score is for Five Persons class (90%) Precision because referring to Fig. 5, we know that TP for Five Persons class is (840) and FP is (96), and the lowest is for Two Persons class Precision (73%) because we know that TP for Two Persons class is (822) and FP is (302). For F1-Score, the highest score is for Five Persons class (86%), and the lowest is for Two Person class (76%). For Recall, the highest score is for One Person class (88%), and the lowest is for Three Persons class (79%). For Precision, the highest score is for Five Persons class (90%), and the lowest is for Two persons classes (73%).

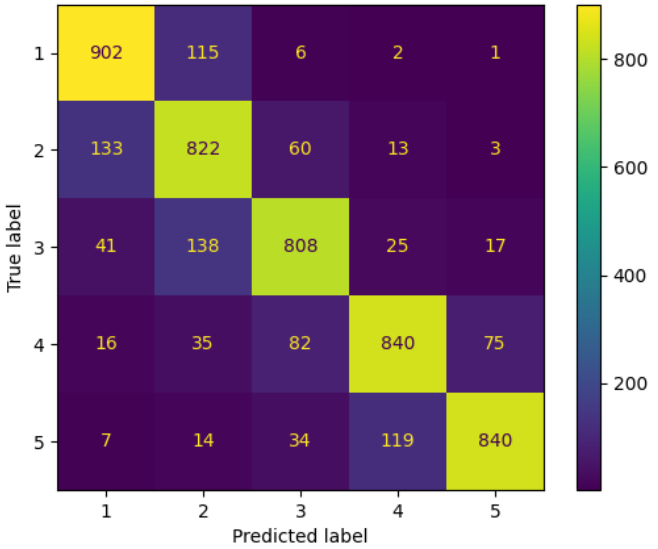

Fig. 5. The Confusion Matrix (RF).

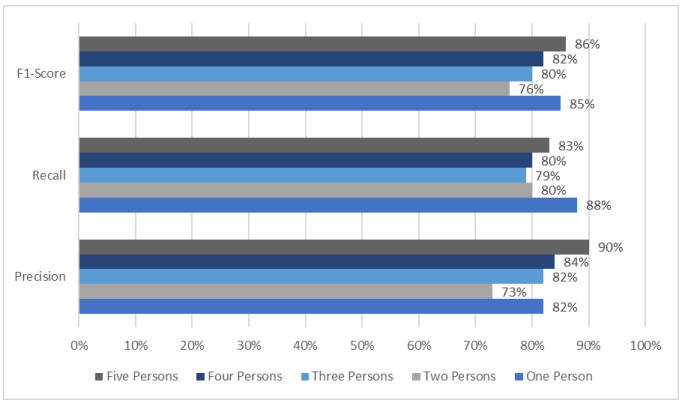

Fig. 6. F1-Score, Recall and Precision (RF).

C. Support Vector Machine (SVM)

Fig. 7 shows the confusion matrix for SVM model. There are five classes. The values vary from the minimum (zero) with purple color to the maximum (972) with dark yellow. The matrix could be read as follows. For Four unknowns, for instance, there are (911) correct predictions, (zero) samples were misclassified as One or Two unknown contributors, (88) samples were misclassified as Three Persons classes and (49) samples were misclassified as Five Persons class. The results show that Five Persons class have the highest number of correct predictions (972), then One Person class with (961), then Three Persons class with (932), then Two Persons class with (918) and finally Four Persons class with (911). In terms of mischaracterization, Four Persons class has the highest number of misclassification (137), then Two Persons class (113), then Three Persons class (97), then One Person class (65) and finally Five Persons class (42).

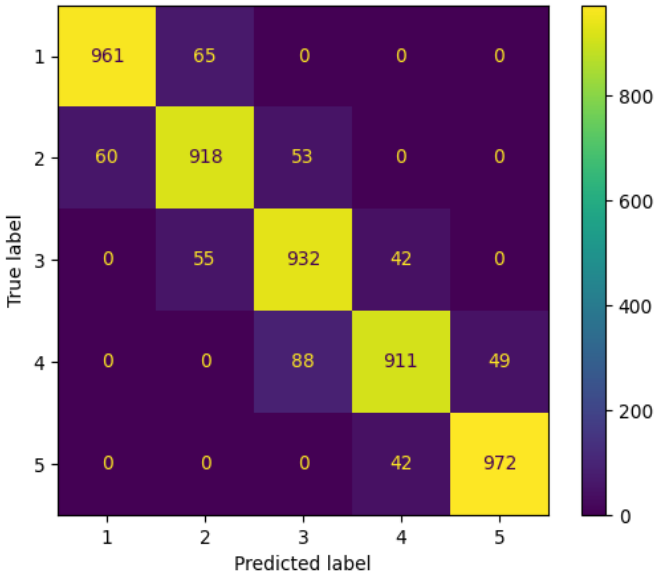

Fig. 7. The Confusion Matrix (SVM).

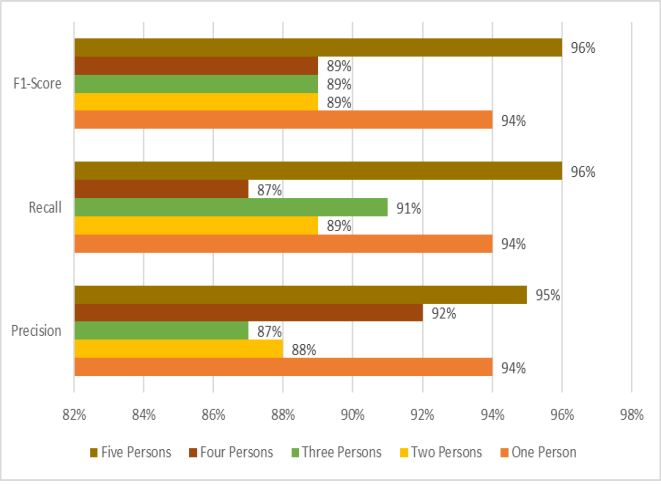

Fig. 8. F1-Score, Recall and Precision (SVM).

Fig. 8 shows F1-Score, Recall and Precision for RF model. The highest score is for both Five Persons class (96%) F1-Score and Five Persons class Recall because referring to Fig. 7 we know that TP for Five Persons class is (972), FP is (49), and FN is (42), and the lowest is for both Four Persons class Recall (87%) and Three Persons class Precision (87%) because we know that TP for Four Persons class is (911) and FN is (137), and TP for Three Persons class is (932), and FP is (141). For F1-Score, the highest score is for Five Persons class (96%), and the lowest is for Four Persons, Three Persons and Two Persons classes (89%). For Recall, the highest score is for Five Persons class (96%), and the lowest is for Four Persons class (87%). For Precision, the highest score is for Five Persons class (95%), and the lowest is for Three persons class (87%).

D. Stochastic Gradient Descent (SGD)

Fig. 9 shows the confusion matrix for SGD model. There are five classes. The values vary from the minimum (zero) with purple color to the maximum (1026) with dark yellow. The matrix could be read as follows. For Five unknowns, for instance, there are (1009) correct predictions, (zero) samples were misclassified as both One or Two unknown contributors, (1) samples were misclassified as Three Persons class and (4) samples were misclassified as Four Persons class. The results show that One Person class have the highest number of correct predictions (1026), then Five Persons class with (1009), then Three Persons class with (748), then Four Persons class with (436) and finally Two Persons class with (118). In terms of mischaracterization, Four Persons class has the highest number of misclassification (612), then Two Persons class (561), then Three Persons class (281), then Five Persons class (5) and finally One Person class (zero).

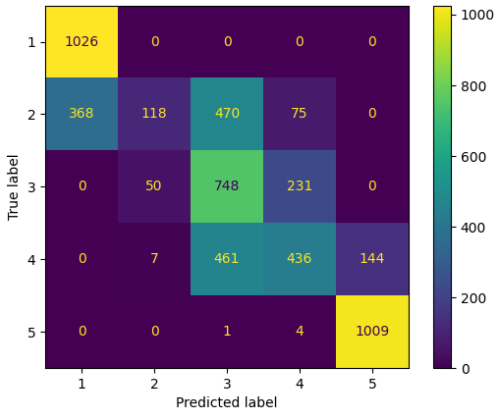

Fig. 9. The Confusion Matrix (SGD).

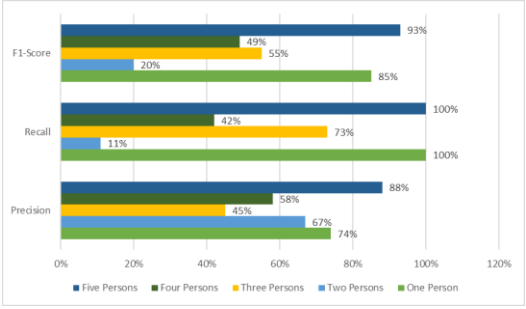

Fig. 10. F1-Score, Recall and Precision (SGD).

Fig. 10 shows F1-Score, Recall and Precision for SGD model. The highest score is for both Five Persons and One Person classes (100%) Recall because referring to Fig. 9, we know that TP for Five Persons class is (1009) and FN is (5), and TP for One Person class is (1026), and FN is (zero), and the lowest is for Two Persons class Precision (11%) because we know that TP for Two Persons class is (118) and FP is (913). For F1-Score, the highest score is for Five Persons class (93%), and the lowest is for Two Persons class (20%). For Recall, the highest score is for both Five Persons and Two Persons classes (100%), and the lowest is for Two Persons class (11%). For Precision, the highest score is for Five Persons class (88%), and the lowest is for Three persons class (45%).

E. Logistic Regression (LR)

Fig. 11 shows the confusion matrix for LR model. There are five classes. The values vary from the minimum (zero) with purple color to the maximum (990) with dark yellow. The matrix could be read as follows. For One number of unknowns, for instance, there are (990) correct predictions, (36) samples were misclassified as Two Persons class, (zero) samples were misclassified as Three, Four or Five unknown contributors. The results show that One Person class have the highest number of correct predictions (990), then Five Persons class with (989), then Three Persons class with (984), then Four Persons class with (958) and finally Two Persons class with (967). In terms of mischaracterization, Four Persons class has the highest number of misclassification (90), then Two Persons class (64), then Three Persons class (45), then Five Persons class (25) and finally One Person class (36).

Fig. 12 shows F1-Score, Recall and Precision for LR model. The highest score is for Five Persons class (98%) Recall because referring to Fig. 11, we know that TP for Five Persons class is (989) and FN is (25), and the lowest is for Four Persons class Recall (91%) because we know that TP for Four Persons class is (958) and FN is (90). For F1-Score, the highest score is for Five Persons class (96%), and the lowest is for Four Persons, Three Persons and Twp Persons classes (94%). For Recall, the highest score is for Five Persons class (98%), and the lowest is for Four Persons class (91%). For Precision, the highest score is for One Person class (97%), and the lowest is for Three persons class (93%).

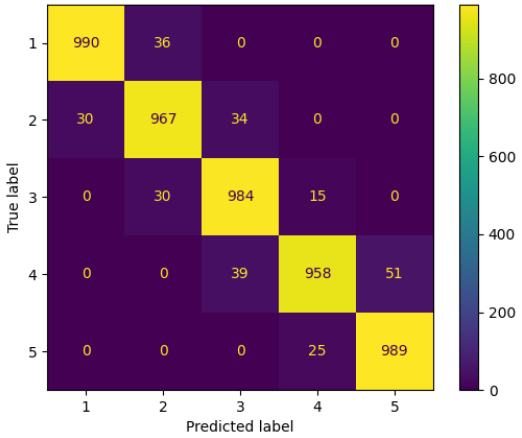

Fig. 11. The Confusion Matrix (LR).

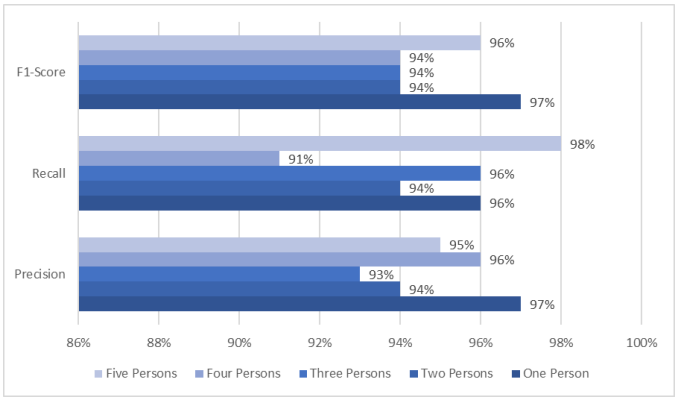

Fig. 12. F1-Score, Recall and Precision (LR).

F. Gaussian NB (GNB)

Fig. 13 shows the confusion matrix for GNB model. There are five classes. The values vary from the minimum (zero) with purple color to the maximum (953) with dark yellow. The matrix could be read as follows. For the Two-Persons class, for instance, there are 772 correct predictions and 259 incorrect predictions. Among these misclassifications, ten samples were misclassified as the One-Person class. Moreover, 231 samples of these were misclassified as the Three-Persons class, 18 samples were misclassified as the Four-Persons class and none of the samples were misclassified as the Five-Persons class. The results show that Three Persons class have the highest number of correct predictions (858), then Two Persons class with (772), then Four Persons class with (760), then Five Persons class with (213) and finally One Person class with (70). In terms of mischaracterization, One Person class has the highest number of misclassification (956), then Five Persons class (801), then Four Persons class (288), then Two Persons class (259) and finally Three Persons class (171).

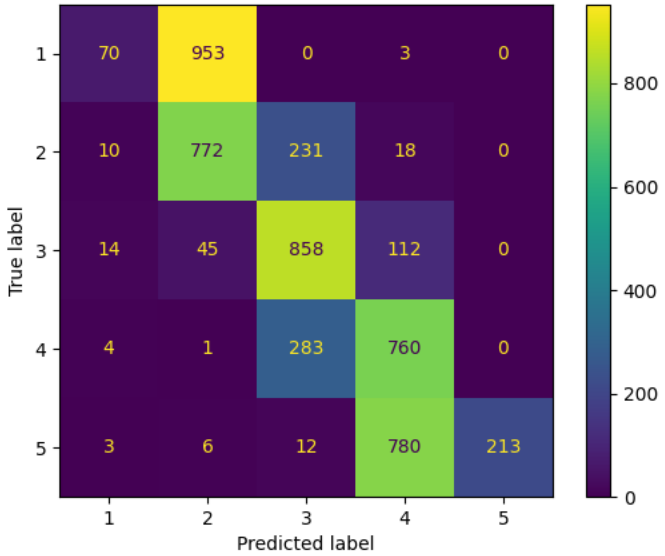

Fig. 13. The Confusion Matrix (GNB).

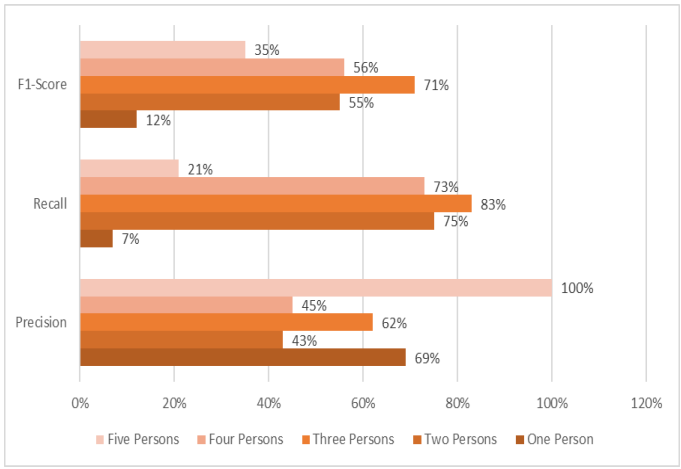

Fig. 14. F1-Score, Recall and Precision (GNB).

Fig. 14 shows F1-Score, Recall and Precision for GNB model. The highest score is for Five Persons class (100%) Precision because referring to Fig. 13, we know that TP for Five-Persons class is (213) and FP is (0), and the lowest is for One-Person class Recall (7%) because we know that TP for One Person class is (70) and FN is (31). For F1-Score, the highest score is for Three Persons class (71%), and the lowest is for One Person class (12%). For Recall, the highest score is for Three Persons class (83%), and the lowest is for One Person class (7%). For Precision, the highest score is for Five Persons class (100%), and the lowest is for Two persons class (43%).

G. Accuracy Comparison

Fig. 15 shows a comparison in terms of Accuracy between the proposed six ML algorithms. The x-axis shows the models names, and the y-axis shows the Accuracy percentage. The results show that LR has the highest score with (95%), then SVM with (91%), then RF with (82%), then SGD with (65%), then KNN with (55%) and finally GNB with (52%).

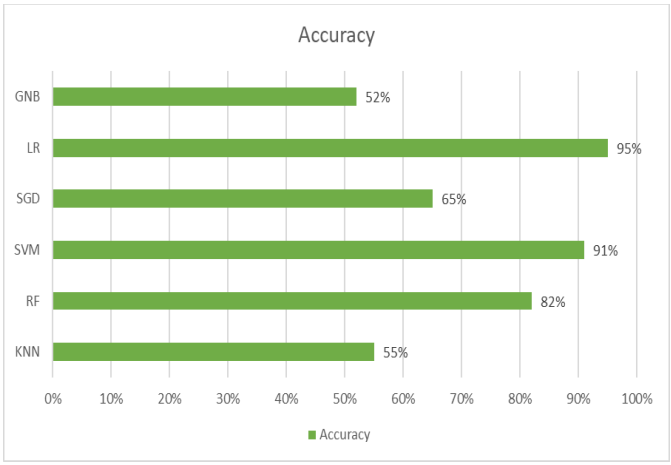

Fig. 15. Accuracy Comparison of the Six ML Algorithms.

H. Comparison with Related Works

Among all the earlier works in the literature on the use of machine or deep learning for estimating the number of unknowns, only Benschop et al. [11] and Kruijver et al. [10] estimated the number of unknowns for DNA mixtures with up to five contributors. The best Accuracy performance for Benschop et al. [11] was reported for the RF algorithm at 83%. The best Accuracy performance for Kruijver et al. [10] was reported for the Decision Trees algorithm at 85%. Comparing these results with our work presented in this paper, we have clearly achieved a better performance, i.e., for the LR algorithm at 95% Accuracy.

V. Conclusions and Future Work

DNA profiling is considered one of the most challenging problems in forensic science. In the near future, the forensic science labs will have more profiles that could have many challenges to deal with, which shows the need for such tools that will help the analysts in their work. Within the next coming years, machine learning will become an essential component in many fields.

This study evaluated six machine learning algorithms with four performance metrics. These are F1-Score, Recall, Precision and Accuracy. The results show that the highest score for KNN is with Five Persons class Precision (75%), the highest score for RF is with Five Persons class Precision (90%), the highest score for SVM is with Five Persons class both F1-Score and Recall (96%), the highest score for SGD is with both Five Persons and One Person class Recall (100%), the highest score for LR is with the Five Persons class Recall (98%), and the highest score for GNB is with Five Persons class Precision (100%). The highest score for F1-Score is with the (LR) 97% One Person class. The highest score for Recall is with the (SGD) 100% One Person class and Five Persons class. The highest score for Precision is with the (GNB) 100% Five Persons class. In terms of Accuracy, the highest score is for the LR with (95%). Comparing with all other related works in the literature, we have clearly achieved a better performance, i.e., for the LR algorithm at 95% Accuracy.

This paper provides an investigation into the performance of machine learning methods for DNA profiling. Further evaluation of machine learning methods is needed and it will form our future work. We will use feature engineering methods to improve the performance of these machine learning methods. We will also investigate tuning the performance of the machine learning methods. Moreover, we will use deep learning to improve classification performance. A major theme of our research is smart cities and societies [23]–[25], big data [26]–[28], high performance computing [29], [30], healthcare [31]–[33], information systems [34], [35], system integration [36], [37], and artificial intelligence [38], [39]. Future work on DNA profiling will also look into developing new smart applications for DNA profiling and its integration with other smart city systems.

References

[1] J. M. Butler, “STR Genotyping and Data Interpretation,” Fundam. Forensic DNA Typing, pp. 205–227, 2010, doi: 10.1016/b978-0-12-374999-4.00010-2.

[2] J. M. Butler, “Applications of DNA Typing,” Fundam. Forensic DNA Typing, pp. 397–421, 2010, doi: 10.1016/b978-0-12-374999-4.00017-5.

[3] J. M. Butler, “Forensic Challenges,” Fundam. Forensic DNA Typing, pp. 315–339, 2010, doi: 10.1016/b978-0-12-374999-4.00014-x.

[4] E. Alamoudi, R. Mehmood, A. Albeshri, and T. Gojobori, “A Survey of Methods and Tools for Large-Scale DNA Mixture Profiling,” EAI/Springer Innov. Commun. Comput., pp. 217–248, 2020, doi: 10.1007/978-3-030-13705-2_9.

[5] E. Alamoudi, R. Mehmood, A. Albeshri, and T. Gojobori, “DNA profiling methods and tools: A review,” in Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, Volume 224, 2018, vol. 224, pp. 216–231, doi: 10.1007/978-3-319-94180-6_22.

[6] T. M. Clayton, J. P. Whitaker, R. Sparkes, and P. Gill, “Analysis and interpretation of mixed forensic stains using DNA STR profiling,” Forensic Sci. Int., vol. 91, no. 1, pp. 55–70, 1998, doi: 10.1016/S0379-0738(97)00175-8.

[7] E. M. Alamoudi, “Parallel Analysis of DNA Profile Mixtures with a Large Number of Contributors,” p. 109, 2019.

[8] M. A. Marciano and J. D. Adelman, “PACE: Probabilistic Assessment for Contributor Estimation— A machine learning-based assessment of the number of contributors in DNA mixtures,” Forensic Sci. Int. Genet., vol. 27, pp. 82–91, Mar. 2017, doi: 10.1016/j.fsigen.2016.11.006.

[9] R. Mehmood, F. Alam, N. N. Albogami, I. Katib, A. Albeshri, and S. M. Altowaijri, “UTiLearn: A Personalised Ubiquitous Teaching and Learning System for Smart Societies,” IEEE Access, vol. 5, pp. 2615–2635, 2017, doi: 10.1109/ACCESS.2017.2668840.

[10] M. Kruijver et al., “Estimating the number of contributors to a DNA profile using decision trees,” Forensic Sci. Int. Genet., vol. 50, no. June 2020, p. 102407, 2021, doi: 10.1016/j.fsigen.2020.102407.

[11] C. C. G. Benschop, J. van der Linden, J. Hoogenboom, R. Ypma, and H. Haned, “Automated estimation of the number of contributors in autosomal short tandem repeat profiles using a machine learning approach,” Forensic Sci. Int. Genet., vol. 43, Nov. 2019, doi: 10.1016/J.FSIGEN.2019.102150.

[12] T. Egeland, I. Dalen, and P. F. Mostad, “Estimating the number of contributors to a DNA profile,” Int. J. Legal Med., vol. 117, no. 5, pp. 271–275, 2003, doi: 10.1007/s00414-003-0382-7.

[13] T. Graversen, “Statistical and Computational Methodology for the Analysis of Forensic DNA Mixtures with Artefacts,” p. 229, 2014.

[14] K. Inman et al., “Lab Retriever: A software tool for calculating likelihood ratios incorporating a probability of drop-out for forensic DNA profiles,” BMC Bioinformatics, vol. 16, no. 1, pp. 1–11, 2015, doi: 10.1186/s12859-015-0740-8.

[15] T. Tvedebrink, P. S. Eriksen, H. S. Mogensen, and N. Morling, “Evaluating the weight of evidence by using quantitative short tandem repeat data in DNA mixtures,” J. R. Stat. Soc. Ser. C Appl. Stat., vol. 59, no. 5, pp. 855–874, 2010, doi: 10.1111/j.1467-9876.2010.00722.x.

[16] Ø. Bleka, G. Storvik, and P. Gill, “EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts,” Forensic Sci. Int. Genet., vol. 21, pp. 35–44, 2016, doi: 10.1016/j.fsigen.2015.11.008.

[17] D. J. Balding, C. D. Steele, D. Building, and G. Street, “likeLTD v6.3: an illustrative analysis, explanation of the model, results of validation tests and version history,” 2016.

[18] H. Swaminathan, C. M. Grgicak, M. Medard, and D. S. Lun, “NOCIt: A computational method to infer the number of contributors to DNA samples analyzed by STR genotyping,” Forensic Sci. Int. Genet., vol. 16, pp. 172–180, 2015, doi: 10.1016/j.fsigen.2014.11.010.

[19] L. E. Alfonse, A. D. Garrett, D. S. Lun, K. R. Duffy, and C. M. Grgicak, “A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt,” Forensic Sci. Int. Genet., vol. 32, no. July 2017, pp. 62–70, 2018, doi: 10.1016/j.fsigen.2017.10.006.

[20] S. Kabiraj et al., “Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm,” 2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020, pp. 11–14, 2020, doi: 10.1109/ICCCNT49239.2020.9225451.

[21] "Stochastic Gradient Descent — Clearly Explained." https://towardsdatascience.com/stochastic-gradient-descent-clearly-explained-53d239905d31.

[22] "Naive Bayes." https://scikit-learn.org/stable/modules/naive_bayes.html.

[23] R. Mehmood, S. See, I. Katib, and I. Chlamtac, Smart Infrastructure and Applications: foundations for smarter cities and societies. Springer International Publishing, Springer Nature Switzerland AG, 2020.

[24] T. Yigitcanlar, L. Butler, E. Windle, K. C. Desouza, R. Mehmood, and J. M. Corchado, "Can Building 'Artificially Intelligent Cities' Safeguard Humanity from Natural Disasters, Pandemics, and Other Catastrophes? An Urban Scholar's Perspective," Sensors, vol. 20, no. 10, p. 2988, May 2020, doi: 10.3390/s20102988.

[25] T. Yigitcanlar, J. M. Corchado, R. Mehmood, R. Y. M. Li, K. Mossberger, and K. Desouza, "Responsible Urban Innovation with Local Government Artificial Intelligence (AI): A Conceptual Framework and Research Agenda," J. Open Innov. Technol. Mark. Complex., vol. 7, no. 1, p. 71, Feb. 2021, doi: 10.3390/joitmc7010071.

[26] Y. Arfat, S. Suma, R. Mehmood, and A. Albeshri, "Parallel shortest path big data graph computations of us road network using apache spark: Survey, architecture, and evaluation," in Smart Infrastructure and Applications Foundations for Smarter Cities and Societies, Springer Cham, 2020, pp. 185–214.

[27] S. Usman, R. Mehmood, and I. Katib, "Big data and hpc convergence for smart infrastructures: A review and proposed architecture," in Smart Infrastructure and Applications Foundations for Smarter Cities and Societies, Springer Cham, 2020, pp. 561–586.

[28] E. Alomari, I. Katib, A. Albeshri, T. Yigitcanlar, R. Mehmood, and A. A. Sa, "Iktishaf+: A Big Data Tool with Automatic Labeling for Road Traffic Social Sensing and Event Detection Using Distributed Machine Learning," Sensors, vol. 21, no. 9, p. 2993, Apr. 2021, doi: 10.3390/s21092993.

[29] S. Alahmadi, T. Mohammed, A. Albeshri, I. Katib, and R. Mehmood, "Performance analysis of sparse matrix-vector multiplication (Spmv) on graphics processing units (gpus)," Electron., vol. 9, no. 10, 2020, doi: 10.3390/electronics9101675.

[30] T. Muhammed, R. Mehmood, A. Albeshri, and I. Katib, "SURAA: A Novel Method and Tool for Loadbalanced and Coalesced SpMV Computations on GPUs," Appl. Sci., vol. 9, no. 5, p. 947, Mar. 2019, doi: 10.3390/app9050947.

[31] T. Muhammed, R. Mehmood, A. Albeshri, and I. Katib, "UbeHealth: A personalized ubiquitous cloud and edge-enabled networked healthcare system for smart cities," IEEE Access, vol. 6, pp. 32258–32285, 2018, doi: 10.1109/ACCESS.2018.2846609.

[32] E. Alomari, I. Katib, A. Albeshri, and R. Mehmood, "Covid-19: Detecting government pandemic measures and public concerns from twitter arabic data using distributed machine learning," Int. J. Environ. Res. Public Health, vol. 18, no. 1, pp. 1–36, Jan. 2021, doi: 10.3390/IJERPH18010282.

[33] S. Alotaibi, R. Mehmood, I. Katib, O. Rana, and A. Albeshri, "Sehaa: A Big Data Analytics Tool for Healthcare Symptoms and Diseases Detection Using Twitter, Apache Spark, and Machine Learning," Appl. Sci., vol. 10, no. 4, p. 1398, Feb. 2020, doi: 10.3390/app10041398.

[34] N. Ahmad and R. Mehmood, "Enterprise systems and performance of future city logistics," Prod. Plan. Control, vol. 27, no. 6, pp. 500–513, Apr. 2016, doi: 10.1080/09537287.2016.1147098.

[35] N. Ahmad and R. Mehmood, "Enterprise systems: Are we ready for future sustainable cities," Supply Chain Manag., vol. 20, no. 3, pp. 264–283, May 2015, doi: 10.1108/SCM-11-2014-0370.

[36] T. Mohammed, A. Albeshri, I. Katib, and R. Mehmood, "UbiPriSEQ—Deep reinforcement learning to manage privacy, security, energy, and QoS in 5G IoT hetnets," Appl. Sci., vol. 10, no. 20, 2020, doi: 10.3390/app10207120.

[37] T. Muhammed, R. Mehmood, A. Albeshri, and A. Alzahrani, "HCDSR: A Hierarchical Clustered Fault Tolerant Routing Technique for IoT-Based Smart Societies," 2020, pp. 609–628.

[38] T. Yigitcanlar et al., "Artificial Intelligence Technologies and Related Urban Planning and Development Concepts: How Are They Perceived and Utilized in Australia?," J. Open Innov. Technol. Mark. Complex., vol. 6, no. 4, p. 187, Dec. 2020, doi: 10.3390/joitmc6040187.

[39] T. Yigitcanlar, R. Mehmood, and J. M. Corchado, "Green Artificial Intelligence: Towards an Efficient, Sustainable and Equitable Technology for Smart Cities and Futures," Sustain. 2021, Vol. 13, Page 8952, vol. 13, no. 16, p. 8952, Aug. 2021, doi: 10.3390/SU13168952.