Applying Explanatory Artificial Intelligence Using LIME to Analyze Arabic Online Handwriting Recognition Models

[1st] Hasanien Ali Talib Alothman[1,2,3]
[1-] *ISITCom - Higher Institute of Computer Science and Communication*
[2-]*LATIS - Laboratory of Advanced Technology and Intelligent Systems, National, Engineering School of Sousse (ENISo), University of Sousse*
Sousse, Tunisia
[3-]*Mechatronics Engineering Department, College of Engineering, University of Mosul*
Mosul, Iraq
hasanien.ali@uomosul.edu.iq

[2nd] Wafa Lejmi
LATIS - Laboratory of Advanced Technology and Intelligent Systems, National, Engineering School of Sousse (ENISo), University of Sousse
Sousse, Tunisia
wafa.lejmi@uc.rnu.tn

[3rd] Safa Ameur
LATIS - Laboratory of Advanced Technology and Intelligent Systems, National, Engineering School of Sousse (ENISo), University of Sousse
Sousse, Tunisia
safaameur@eniso.u-sousse.tn

[4th] Mohamed Ali Mahjoub
LATIS - Laboratory of Advanced Technology and Intelligent Systems, National, Engineering School of Sousse (ENISo), University of Sousse
Sousse, Tunisia
mohamedali.mahjoub@eniso.rnu.tn

***Abstract*—This study applies interpretable artificial intelligence (XAI) techniques in an unprecedented way to improve transparency in online Arabic handwriting recognition models. By using the LIME (Local Model-Independent Explanations) technique to analyze and interpret a convolutional neural network (CNN) model trained on the ADAB dataset. LIME highlights the key features that the model relies on in handwriting recognition. This research demonstrates the effectiveness of the LIME technique in uncovering the factors that the model relies on in an unconventional way and provides detailed explanations of how the decisions are made, which enhances transparency and increases users' confidence in these models. The results provide useful information for improving the performance of handwriting recognition systems and a beginning for applying XAI techniques on a larger scale, including using more complex datasets and additional interpretation techniques to gain a deeper understanding of how models work with different handwriting styles. This study paves the way for the use of interpretable techniques in the field of Arabic handwriting recognition.**

Keywords— Arabic online handwriting, XAI, LIME, CNN, interpretability, deep learning.

The rapid developments in artificial intelligence and the great increase in the use of deep learning techniques have led to the need for a method that explains the reasons for the decisions taken by these models [1], which are considered black boxes that are difficult to know the basis on which they relied on in their results due to the complexity of the computational processes within them [2]. There is a widespread consensus that transparency of the process is necessary, especially for end users, data owners and affected entities when using applications and services based on artificial intelligence. All of this led to the emergence of the concept of explainable artificial intelligence prominently in the late second decade of the twenty-first century [3]. Many research and government institutions launched initiatives to support research in the field of XAI (eXplainable Artificial Intelligence), such as DARPA (Defense Advanced Research Projects Agency), which launched the Explainable AI initiative, which aims to develop artificial intelligence systems that can easily explain their decisions [4]. There are several important reasons for the emergence of XAI:

A. Increasing complexity

It has become difficult to elucidate how cognitive function models such as deep learning work, and how they make decisions [2]. For example, model that works to distinguish between cats and dogs. XAI techniques break down patterns in certain features in the model such as shapes of eyes or shapes of ears, which clarify how the model works [3].

B. Sensitive applications

Artificial intelligence has entered the application field in many sensitive fields that require transparency behind the reasons for making decisions [3]. For example, if an AI model is used in medical fields to diagnose a specific disease such as cancer, the medical staff must understand why the model recommended a specific treatment plan. The model may recommend amputation of a body part based on a wrong diagnosis if these explanations are not available [5]. As a result, the rationale behind XAI becomes essential in such settings to guarantee that the model is rooted in transparent, logical data and based on scientific principles [6].

C. Responsibility and accountability

When making critical decisions based on AI, officials must be able to explain how the AI-powered system reached these decisions to avoid legal liability and errors [6] [3]. For example, in self-driving cars, if an accident occurs, the self-driving system must be able to explain how it made the decisions that led to this accident to indicate whether the system caused this accident or not and the extent to which its decisions were correct. XAI has the ability to generate explanations such as a car suddenly stopping to avoid an accident due to a misread red light or identifying a specific object as a danger [4].

D. Laws and regulations

There are laws that require clarification of the decisions made by AI-powered systems [2] [7]. A good instance is within the European Union and the General Data Protection Regulation (GDPR), according to which people are entitled to being aware of how decisions impacting them are reached,

such as decisions related to credit or employment that rely on AI-powered applications [1].

E. Improving models

The ability of XAI systems to detect errors and biases in AI models allows researchers and developers to improve them [4]. For instance, an AI model used in financial fraud detection systems can be very good at detecting fraudulent activities; however, when its decisions are interpreted using XAI, software developers may find out that it relies on biased patterns in making decisions, like prioritizing certain categories/geographical locations unfairly over the rest [3].

F. Interaction between humans and machines

Interpreting AI decisions facilitates the interaction between humans and AI-powered systems and increases their capabilities and effectiveness [4]. Users tend to trust models that provide clear explanations more than those that don't [1].

In the recent past, there has been rampant use of XAI technologies, and commensurably much focus has been put on developing techniques like LIME, SHAP, and Grad-CAM that help AI practitioners get a deep insight into what goes on inside models [3], [5]. Specifically, LIME provides to relatively complex decisions local explanations [3] by assessing the effect of each feature on the prediction and therefore reveals the key reason why the model reached a certain decision based on particular features [8].

In this context, the importance of XAI in the field of recognizing live writing in Arabic emerges, which is one of the major challenges due to the complexities of the language from the diversity of fonts and formations and the change in the shape of letters based on their position in the word [9]. This chapter delves into the use of such approaches, specifically LIME, to interpret results and improve the effectiveness of the Arabic handwriting recognition model. This enables us to understand how Arabic handwriting recognition models arrive at decisions by interpreting the most salient temporal and spatial characteristics.

II. Background on XAI Techniques

Here are listed the most prominent XAI techniques that have proven successful in explaining the complex decisions made by AI models. These techniques or approaches depend on offering profound insights into ways in which various characteristics influence model outcomes, thus boosting clarity and boosting trust in them. In addition, these techniques enable the detection of potential errors, biases, and help improve the overall performance of models [10]. Below we list the XAI techniques that have proven effective in practical applications, and at the end of this paragraph, table I is attached, explaining the differences between them in terms of advantages, disadvantages, and best applications.

A. LIME (Local Interpretable Model-agnostic Explanations)

LIME is a technique used to provide local explanations for predictions made by complex models [11]. LIME changes the input data somewhat in order to discover how the model output changes. This helps to understand how certain features affect the predictions made by the model [12]. LIME relies on creating a local simplified model around each prediction to understand how the model arrived at its decision [13]. LIME creates a simplified local version of the model (such as a simple linear model) to explain how the features affected the prediction in a particular case. This is applied to each prediction separately, allowing for a more granular understanding of how the model made the decision [14].

For example, in fig.1, LIME is used to explain how an AI model made its decision [15], that an image contains a "Labrador Retriever". On the left, you can see the original image containing a dog and a cat, and the model correctly classified the dog as a "Labrador Retriever". The question LIME is trying to answer is: Why did the model classify this dog this way?. On the right side of the image, LIME shows the regions that were influential in the model's decision. As you can see, LIME highlighted parts of the dog's face as the most important areas that contributed to the prediction that the dog was a "Labrador." These colored areas on the image indicate the features that the model relied on to make its decision. In this example, LIME does this by analyzing that image, dividing the image into multiple regions, adjusting some parts then determining how much each part of the image played a role in the final outcome [16]. In this case, LIME shows that the head or ear shape was the main reason the model decided that the image contained a "Labrador," rather than, for example, any other type of dog or animal in the image, such as a cat.

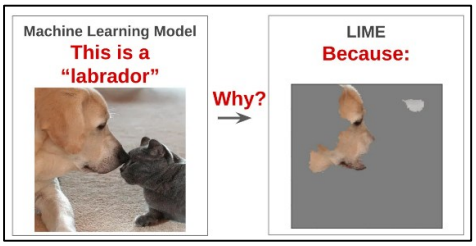

Fig. 1. Local influence map showing the regions the model relied on to classify the image as a dog 'Labrador Retriever' using LIME. [15]

B. SHAP (SHapley Additive exPlanations)

A game theory technique for providing explanations for each feature in the model [17]. SHAP works by assigning a "Shapley value" to each feature, which represents its contribution to the final prediction [18]. This method is effective in providing a comprehensive explanation of the global and local effects on the model's output. This value indicates the contribution of each feature to the final decision taken by the model, thereby ensuring both comprehensive global effects across all predictions and local effects for particular instances [19]. The aim of the SHAP is to share the gain equally among the players in accordance with their contribution to an outcome. A player in this particular outcome model can be thought of as a "feature," and it is a prediction of how such a "feature" has at the end influenced the outcome [20].

For example, in [21], of SHAP being used in a diabetes risk prediction model, we can distinguish between the global and local effects of each feature,. As shown in fig. 2, when looking at an individual case of a specific person aged 65 years and whose blood pressure is 180, SHAP explains exactly how these features affected the final outcome. In this case, age had the greatest impact (+0.4) in raising the score to 0.4. Gender also contributed (+0.1) to the score. While the contributions of blood pressure and BMI were less significant. Thus, a global explanation explains the overall impact of features on the model as a whole, while a local explanation focuses on explaining the individual predictions for specific cases.

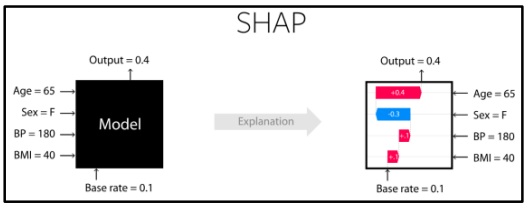

Fig. 2. SHAP interpretation diagram of local and global contributions and influences on the model[21].

C. LRP (Layered Relevance Propagation)

This is a methodology for forming explanations that leverage the stratifications of deep neural networks (DNNs) [22]. By providing a feature map that pertains to the predictions, one can gain insight into what is relevant and, therefore, understand how a given layer of the neural network affects the last decision [23]. LRP moves the influence of the input features across the layers of the neural network and shows how each layer and each feature contributes to the final prediction. This technique is used to demonstrate how each of the characteristics influences the last forecast and then creates diagrams of the heat to depict these impacts [24].

D. Heatmap

The heat map is a visualization method which allows depicting the regions of interest for the model in images. This technique is most often used in conjunction with neural networks, and in particular, convolutional ones, to figure out the classification of images in order to make a decision [4].

E. SEDC and SEDC-T

SEDC (Sustained Feature Exclusion Explanation) and SEDC-T (SEDC-Temporal) are two methods employed in feature removal-based interpretation to understand the impact on the model's decision when certain features are removed., and if removed, the decision would change [25].

F. Feature Importance

Feature Importance is a technique used to determine how much impact each feature has on the final prediction [25]. It is frequently employed in linear models like logistic regression and non-linear ones like random forests [26]. The significance or magnitude of every feature engineered into the model is calculated to decide upon its relevance in decision making.

For this instance, an importance map in a model for classifying cancer depending on diverse features may indicate that in "tumor size" and "incidence" parameters are very relevant to the end results [27].

G. Induction

The model may use decision trees or rules to provide explanations on how it arrived at a certain decision by taking into account the input features and therefore giving an explanation that is clear and rational [28].

H. Provenance (Data Origin)

In AI, from where decisions are made, a tree was formed; either in form of decision trees or rules in which explanations were given through input characters thus forming a rational and transparent interpretation [29]. For instance when we are using a decision tree to sort individuals according to their monetary status we can use it as an example because the most important features that led to such classification were annual salaries and working backgrounds [30].

TABLE I. COMPARATIVE TABLE OF XAI TECHNOLOGIES

Technology	Features	Defects	Future Directions
LIME	Instant usability Local interpretation is easy to present.	Explanations may be unstable The ranking does not take into account the dependence between features.	Reduce local instability by improving repetition-based explanations.
SHAP	Improved speed. It provides a global view of interpretations.	Small modifications can lead to drastic changes in interpretation.	Use SHAP to define the individual contribution of each feature more precisely.
LRP	Suitable for neural networks. Provides a link map relevant to forecasts.	Interpretation is at a low abstract level.	Improve stratified interpretations of neural networks.
Heatmap	Visual display depends on the importance of features.	Individual pixels do not provide clear meaning to humans.	Use this technique in classifying text next to images.
SEDC and SEDC-T	Explanations focused on the human element.	It may offer more than one irreducible explanation.	The use of reverse analysis of facts in the classification of texts.
Feature Importance	An interpretation based on feature weights.	Limited to local features only, which may divert attention from vital global one.	Reduce local influence to improve forecasts.
Induction	convenient for programming. It is based on logical & arboreal rules.	It requires deep understanding from the end user.	Generalization is required for miscellaneous data.
Provenance	Interpretations based on natural language make accessible to human use.	Not fully compatible with mathematical verification as it relies on natural language.	Develop verification mechanisms to increase reliability.

III. POSSIBLE USE OF XAI IN ARABIC HANDWRITING RECOGNITION

XAI techniques are crucial for understanding decision-making systems used in AI [3]. In the domain of handwriting recognition specifically, it is required to know which aspects the model depends upon when making its choice [31]. The concept is similar to that of reinforcing the specificness of the models used in model selection and improving the accountability [32] which is very much required when working with applications like hand writing recognition.

A. The importance of XAI in the field of handwriting recognition

The difficulty in explaining and understanding the decisions that machine learning models make is one of the components in the high acceptance of the assumption that machine learning models act as black boxes in applications of artificial intelligence [33]. This is why, there is a high recognition for XAI, as the processes within the model can be understood in terms of exposing the most important features for the prediction. In the field of Arabic handwriting recognition, XAI can be used to explain:

- Why the machine learning model classified a certain letter or word in a certain way [34].
- How the model reached its decision, and what features it relied on most [29].

XAI can help understand how the model interacts with specific elements of text, such as:

- Changes in the shapes between letters due to their position in the word.
- Variations in the spacing between letters that may affect recognition accuracy.
- The curvatures and bends of Arabic letters that may vary significantly between people's handwriting.

B. *How to apply XAI to Arabic handwriting recognition*

We can apply XAI techniques such as LIME to an Arabic handwriting recognition model to understand how the model relies on certain features to make its decisions. Using LIME, we can extract and interpret the features that the model considers most influential when recognizing written words. For example, the writings are analyzed using convolutional models (such as CNN), and then LIME is applied to highlight the important parts of the image that prompted the model to make a particular decision.

C. *Advantages of using XAI to improve predictive models*

Introducing XAI into handwriting recognition models brings many advantages, such as:

- Increased transparency: Users can see how and why the model made its decision, which improves users' confidence in the systems [35].
- Improving performance: By identifying the points on which the model is incorrectly relying, these points can be modified and improved to enhance the model's performance [23].
- Improvements: XAI helps identify errors that the model may make, such as errors resulting from irrelevant features that may affect the prediction accuracy [36].

D. *Potential challenges when using XAI in handwriting recognition*

Despite the great benefits of applying XAI in this field, there are challenges that must be taken into account, such as the great diversity in handwriting styles vary greatly between people [37], making it difficult for XAI to identify uniform features for everyone. The complexity of the Arabic language: The unique nature of Arabic letters in terms of their connection to each other [38] may require specialized methods to analyze features effectively. Lastly, the lack of data: there is limited amount of handwriting data available in Arabic compared to other languages [31], which limits the performance of machine learning models.

IV. Applying LIME to an Arabic Handwriting Recognition Model

The approach of using LIME was adopted in this study because it provides localized explanations that assist in interpreting the rationale of the model in cases that require complex reasoning, such as Arab handwriting recognition. LIME is also appropriate for explaining complex models, especially neural networks, which are the main focus in this field since it explains the effect of various features on the model outcome.

A. *CNN model*

This model contains the following sequence of layers:

- An input layer to accept grayscale 100x100 pixel images (single channel).
- Two Conv2D layers with 32 3x3 filters each, succeeded by a MaxPooling2D layer for decreasing spatial dimensions.The flattening layer changes the 2D matrix to a single vector to use in the dense layers.
- Two dense layers; 128 units and 64 units to extract high-dimensional features.
- The output layer acts as a Softmax layer with 7 units representing 7 different handwriting classes.
- Since the input was high-dimensional, 32 filters were utilized.
- When the pixel density was increased from 28x28 pixels to 100x100 pixels, more detailed information became available.
- The complexity of the features was increased by expanding the densely connected layers so as to enable more complex patterns be recognized in dataset.

B. *Training and evaluation of the model*

The processed data was used to train the model. For training and testing, this data was split into 60% and 20%, respectively. A loss was calculated by a categorical crossentropy function, while the Adams optimizer was used to adjust the learning rate. A number of performance metrics, such as specific recall and the F1 coefficient, were utilized in evaluating the model's performance. Furthermore, as depicted in fig.3, a confusion matrix was generated to show how well the model can differentiate between different handwriting classes. As indicated by the table below, accuracy on the given dataset was 97.14%.

TABLE II. Performance Metrics of the Proposed CNN Model for Arabic Handwriting Recognition.

Metric	Value (%)
Accuracy	97.14%
Precision	97.62%
Recall	97.14%
F1-Score	97.11%

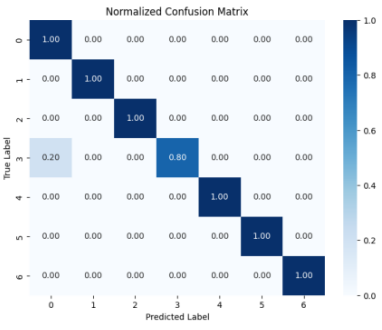

Fig. 3. Confusion matrix of the proposed model.

V. How Does LIME Interpret the CNN Model

In this part, we will discuss how the recognition of handwritten Arabic characters, which is performed using

CNN, relies on LIME to explain its decisions. The explanation is a demonstration of how LIME effectively gives local explanations into the model decisions through mathematical methods and algorithmic application. LIME mathematical formulations and step by step methods to implement are discussed in the next sub-section.

A. Mathematical Equations for LIME

LIME works by creating a simple local model around each data point to explain the decision-making process of the original model. Thus, the essence of LIME is to minimize the deviation between the output of the complex initial model and the simple local model $g(z')$ using a loss function, eq:

$$L(f, g, \pi_x) = \sum_{z' \in Z} \pi_x(z')(f(z') - g(z'))^2 \qquad (1)$$

B. Steps of the LIME Algorithm

Algorithm.: LIME for CNN Handwriting Recognition
Require: Image input *X*
Ensure: Local explanation for class prediction *Z*

1-Input Preparation:

$$\textbf{For } i = 1 \textbf{ to } n: \quad X_i \leftarrow convert\_to\_image(X_i)$$

2-Apply image reshaping and normalization to match CNN input dimensions.

$$P_{CNN}(X') = Softmax(W_{CNN}CNN \cdot X' + b_{CNN})$$

Where $X' \in \mathbf{R}^{100\times100}$ and W_{CNN} are CNN Weight.

3-LIME Sampling:

Select a number of perturbed samples $X'_{perturbed}$ for local interpretation:

$$X'_{perturbed} \leftarrow perturb(X')$$

This step generates perturbed images by hiding or modifying parts of the original image.

4-Local Explanation Model:

Train a local interpretable model $g(z')$ using linear regression for the perturbed samples:

$$g(z') = W_g \cdot z' + b_g$$

z' is the perturbed sample and $g(z')$ is the output of the local explanation model.

5-Weight Assignment for Locality:

For each perturbed sample, assign a weight $\pi_x(z')$ based on its proximity to the original sample **X**:

$$\pi x(z') = exp\left(-\frac{d(X, z')}{\sigma^2}\right)$$

$d(X, z')$is the distance between the original and perturbed sample.

6-Explanation Generation:

Minimize the loss function between the original model's prediction:

$$Loss = \sum_{z' \in Z} \pi_x(z')(P_{CNN}(z') - g(z'))^2$$

7-Identify Features:

Identify the top k features (image segments) that have the highest impact on the local explanation:

$$Top\_Features = arg\ max\ W_g[i]\ Where\ i = 1, \ldots, k$$

8-Visualization:

Display the original image and the perturbed regions used to explain the model decision:

$$Z \leftarrow mark\_boundaries(X', Top\_Features)$$

9-Return:

Return the visual explanation *Z*

VI. Experimental Results:

The experiments were conducted using Google Colab Pro with a T4 GPU and Python 3.10.12 in the TensorFlow and Keras environment. The CNN model was trained on the ADAB dataset, aiming to evaluate the model's ability to recognize handwritten Arabic letters and use LIME for interpreting decisions. LIME highlighted influential features in each image: green areas contributed positively, red areas negatively, and yellow borders marked key regions. For the word "شلوسة (fig. 4), LIME underscored the curved features that aided classification. In another example (fig. 5), the model struggled to distinguish between classes 0 and 3 by 20%, as seen in the confusion matrix. LIME provided transparency in model decisions, enhancing confidence in using XAI techniques for Arabic handwriting recognition.

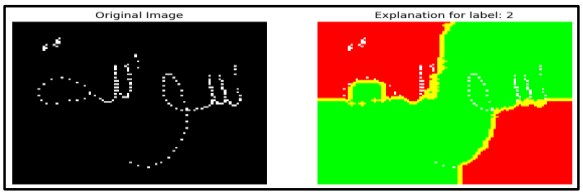

Fig. 4. Image & LIME Explanation for Arabic Handwriting Label: 2

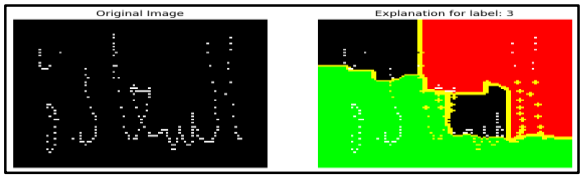

Fig. 5. Image and LIME Explanation for Arabic Handwriting Label: 3

VII. Conclusion

In this research, an unprecedented attempt is made to apply XAI via LIME to contribute to Arabic handwriting recognition, and the results show encouraging promise. The LIME algorithm helped generate local interpretations that highlight the most important features on which the model made its decisions. In the course of the research, LIME proved its usefulness by facilitating the identification of key portions of the text that influenced the prediction, thus enhancing the interpretability and transparency of the model. This is particularly relevant to the recognition of Arabic handwriting as the shapes of the letters morph depending on their position within the word itself.

The incorporation of LIME within a CNN architecture provides insights into the decision-making process of the model while classifying complicated data. This improvement serves to increase the trustworthiness of AI driven models.

These results open the way to using a larger and more diverse dataset to experiment with the technology on a larger scale. In addition, LIME can be applied at the level of individual letters for finer analysis of features, or focus on the temporal properties of Arabic handwriting for a deeper understanding of patterns. Other XAI techniques, such as SHAP or Grad-CAM, could also be tried, making this area open for many future researches into improving handwriting recognition models and making them more transparent and interpretable.

References

[1] I. Ahmed, G. Jeon, and F. Piccialli, 'From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where', *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5031–5042, 2022, doi: 10.1109/TII.2022.3146552.

[2] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, 'Benchmarking and survey of explanation methods for black box models', *Data Mining and Knowledge Discovery*, vol. 37, no. 5, pp. 1719–1778, Sep. 2023, doi: 10.1007/s10618-023-00933-9.

[3] A. R. Javed, W. Ahmed, S. Pandya, P. K. R. Maddikunta, M. Alazab, and T. R. Gadekallu, 'A Survey of Explainable Artificial

Intelligence for Smart Cities', *Electronics*, vol. 12, no. 4, 2023, doi: 10.3390/electronics12041020.
[4] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, 'A historical perspective of explainable Artificial Intelligence', *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 1, p. e1391, 2021, doi: https://doi.org/10.1002/widm.1391.
[5] R. Dazeley, P. Vamplew, C. Foale, C. Young, S. Aryal, and F. Cruz, 'Levels of explainable artificial intelligence for human-aligned conversational explanations', *Artificial Intelligence*, vol. 299, p. 103525, Oct. 2021, doi: 10.1016/j.artint.2021.103525.
[6] G. Vilone and L. Longo, 'Classification of Explainable Artificial Intelligence Methods through Their Output Formats', *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 615–661, 2021, doi: 10.3390/make3030032.
[7] A. Das and P. Rad, 'Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey'. 2020. [Online]. Available: https://arxiv.org/abs/2006.11371
[8] M. Nazar, M. M. Alam, E. Yafi, and M. M. Su'ud, 'A Systematic Review of Human–Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques', *IEEE Access*, vol. 9, pp. 153316–153348, 2021, doi: 10.1109/ACCESS.2021.3127881.
[9] H. A. T. Alothman, W. Lejmi, and M. A. Mahjoub, 'New Approach Based on Substantial Derivative and LSTM for Online Arabic Handwriting Script Recognition', in *Proceedings of the 16th International Conference on Agents and Artificial Intelligence, ICAART 2024*, A. P. Rocha, L. Steels, and H. J. van den Herik, Eds., Rome, Italy: SCITEPRESS, 2024, pp. 689–698. doi: 10.5220/0012385000003636.
[10] O. O. Olateju, S. U. Okon, O. O. Olaniyi, A. D. Samuel-Okon, and C. U. Asonze, 'Exploring the Concept of Explainable AI and Developing Information Governance Standards for Enhancing Trust and Transparency in Handling Customer Data', *Journal of Engineering Research and Reports*, vol. 26, no. 7, pp. 244–268, Jun. 2024, doi: 10.9734/jerr/2024/v26i71206.
[11] E. Amparore, A. Perotti, and P. Bajardi, 'To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods', *PeerJ Computer Science*, vol. 7, p. e479, Apr. 2021, doi: 10.7717/peerj-cs.479.
[12] S. Ahmed, M. Shamim Kaiser, M. S. Hossain, and K. Andersson, 'A Comparative Analysis of LIME and SHAP Interpreters with Explainable ML-Based Diabetes Predictions', *IEEE Access*, pp. 1–1, 2024, doi: 10.1109/ACCESS.2024.3422319.
[13] V. Vimbi, N. Shaffi, and M. Mahmud, 'Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection', *Brain Informatics*, vol. 11, no. 1, p. 10, Apr. 2024, doi: 10.1186/s40708-024-00222-1.
[14] Z. Tan, Y. Tian, and J. Li, 'GLIME: General, Stable and Local LIME Explanation'. 2023. [Online]. Available: https://arxiv.org/abs/2311.15722
[15] C. Arteaga, 'Interpretable Machine Learning for Image Classification with LIME', Medium. Accessed: Sep. 27, 2024. [Online]. Available: https://towardsdatascience.com/interpretable-machine-learning-for-image-classification-with-lime-ea947e82ca13
[16] A. H. Oveis, E. Giusti, S. Ghio, G. Meucci, and M. Martorella, 'LIME-Assisted Automatic Target Recognition With SAR Images: Toward Incremental Learning and Explainability', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 9175–9192, 2023, doi: 10.1109/JSTARS.2023.3318675.
[17] Y. Gebreyesus, D. Dalton, S. Nixon, D. De Chiara, and M. Chinnici, 'Machine Learning for Data Center Optimizations: Feature Selection Using Shapley Additive exPlanation (SHAP)', *Future Internet*, vol. 15, no. 3, 2023, doi: 10.3390/fi15030088.
[18] D. Fryer, I. Strümke, and H. Nguyen, 'Shapley values for feature selection: The good, the bad, and the axioms', arXiv.org. Accessed: Sep. 27, 2024. [Online]. Available: https://arxiv.org/abs/2102.10936v1
[19] Z. Li, 'Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost', *Computers, Environment and Urban Systems*, vol. 96, p. 101845, 2022, doi: https://doi.org/10.1016/j.compenvurbsys.2022.101845.
[20] H. Chen, I. C. Covert, S. M. Lundberg, and S.-I. Lee, 'Algorithms to estimate Shapley value feature attributions', *Nature Machine Intelligence*, vol. 5, no. 6, pp. 590–601, Jun. 2023, doi: 10.1038/s42256-023-00657-x.
[21] S. Lundberg, 'SHAP: A game theoretic approach to explain the output of machine learning models'. 2018. [Online]. Available: https://github.com/shap/shap
[22] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, 'Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications', *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021, doi: 10.1109/JPROC.2021.3060483.
[23] L. Weber, S. Lapuschkin, A. Binder, and W. Samek, 'Beyond explaining: Opportunities and challenges of XAI-based model improvement', *Information Fusion*, vol. 92, pp. 154–176, 2023, doi: https://doi.org/10.1016/j.inffus.2022.11.013.
[24] R. Guerrero-Gómez-Olmedo, J. L. Salmeron, and C. Kuchkovsky, 'LRP-Based path relevances for global explanation of deep architectures', *Neurocomputing*, vol. 381, pp. 252–260, 2020, doi: https://doi.org/10.1016/j.neucom.2019.11.059.
[25] G. Elkhawaga, O. Elzeki, M. Abuelkheir, and M. Reichert, 'Evaluating Explainable Artificial Intelligence Methods Based on Feature Elimination: A Functionality-Grounded Approach', *Electronics*, vol. 12, no. 7, 2023, doi: 10.3390/electronics12071670.
[26] S. M. Kasongo and Y. Sun, 'Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset', *Journal of Big Data*, vol. 7, no. 1, p. 105, Nov. 2020, doi: 10.1186/s40537-020-00379-6.
[27] G. Shoham, A. Berl, O. Shir-Az, S. Shabo, and A. Shalom, 'Predicting Mohs surgery complexity by applying machine learning to patient demographics and tumor characteristics.', *Exp Dermatol*, vol. 31, no. 7, pp. 1029–1035, Jul. 2022, doi: 10.1111/exd.14550.
[28] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, 'Explanation of Machine Learning Models Using Improved Shapley Additive Explanation', in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, in BCB '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 546. doi: 10.1145/3307339.3343255.
[29] R. Dwivedi *et al.*, 'Explainable AI (XAI): Core Ideas, Techniques, and Solutions', *ACM Comput. Surv.*, vol. 55, no. 9, Jan. 2023, doi: 10.1145/3561048.
[30] C. Sampaio, 'Get Feature Importances for Random Forest with Python and Scikit-Learn', Stack Abuse. Accessed: Sep. 28, 2024. [Online]. Available: https://stackabuse.com/get-feature-importances-for-random-forests-with-python-and-scikit-learn/
[31] M. Eltay, A. Zidouri, and I. Ahmad, 'Exploring Deep Learning Approaches to Recognize Handwritten Arabic Texts', *IEEE Access*, vol. 8, pp. 89882–89898, 2020, doi: 10.1109/ACCESS.2020.2994248.
[32] B. Kim, J. Park, and J. Suh, 'Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information', *Decision Support Systems*, vol. 134, p. 113302, 2020, doi: https://doi.org/10.1016/j.dss.2020.113302.
[33] B. Brożek, M. Furman, M. Jakubiec, and B. Kucharzyk, 'The black box problem revisited. Real and imaginary challenges for automated legal decision making', *Artificial Intelligence and Law*, vol. 32, no. 2, pp. 427–440, Jun. 2024, doi: 10.1007/s10506-023-09356-9.
[34] S. Ali *et al.*, 'Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence', *Information Fusion*, vol. 99, p. 101805, 2023, doi: https://doi.org/10.1016/j.inffus.2023.101805.
[35] W. J. von Eschenbach, 'Transparency and the Black Box Problem: Why We Do Not Trust AI', *Philosophy & Technology*, vol. 34, no. 4, pp. 1607–1622, Dec. 2021, doi: 10.1007/s13347-021-00477-0.
[36] W. Saeed and C. Omlin, 'Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities', *Knowledge-Based Systems*, vol. 263, p. 110273, 2023, doi: https://doi.org/10.1016/j.knosys.2023.110273.
[37] A. Bin Durayhim, A. Al-Ajlan, I. Al-Turaiki, and N. Altwaijry, 'Towards Accurate Children's Arabic Handwriting Recognition via Deep Learning', *Applied Sciences*, vol. 13, no. 3, Art. no. 3, Jan. 2023, doi: 10.3390/app13031692.
[38] H. A. T. Alothman, W. Lejmi, and M. A. Mahjoub, 'A Survey on the Online Arabic Handwriting Recognition: Challenges, Datasets and Future Directions', in *17th International Conference on Machine Vision (ICMV 2024)*, W. Osten, Ed., SPIE, 2024.