IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 13 April 2023, accepted 5 May 2023, date of publication 10 May 2023, date of current version 18 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3274851

RESEARCH ARTICLE

SignExplainer: An Explainable AI-Enabled Framework for Sign Language Recognition With Ensemble Learning

DEEP R. KOTHADIYA[1,2], CHINTAN M. BHATT[3], (Member, IEEE), AMJAD REHMAN[2], (Senior Member, IEEE), FATEN S. ALAMRI[4], AND TANZILA SABA[2], (Senior Member, IEEE)

[1]U & P U Patel Department of Computer Engineering, Faculty of Technology (FTE), Chandubhai S. Patel Institute of Technology (CSPIT), Charotar University of Science and Technology (CHARUSAT), Changa 388421, India
[2]Artificial Intelligence and Data Analytics Laboratory (AIDA), College of Computer and Information Sciences (CCIS), Prince Sultan University, Riyadh 11586, Saudi Arabia
[3]Department of Computer Science and Engineering, School of Engineering and Technology, Pandit Deendayal Energy University, Gandhinagar, Gujarat 382007, India
[4]Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia

Corresponding authors: Deep R. Kothadiya (deepkothadiya.ce@charusat.ac.in) and Chintan M. Bhatt (chintan.bhatt@sot.pdpu.ac.in)

This work was supported by Princess Nourah bint Abdulrahman University Researchers Supporting through Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, under Project PNURSP2023R346.

ABSTRACT Deep learning has significantly aided current advancements in artificial intelligence. Deep learning techniques have significantly outperformed more than typical machine learning approaches, in various fields like Computer Vision, Natural Language Processing (NLP), Robotics Science, and Human-Computer Interaction (HCI). Deep learning models are ineffective in outlining their fundamental mechanism. That's the reason the deep learning model mainly consider as Black-Box. To establish confidence and responsibility, deep learning applications need to explain the model's decision in addition to the prediction of results. The explainable AI (XAI) research has created methods that offer these interpretations for already trained neural networks. It's highly recommended for computer vision tasks relevant to medical science, defense system, and many more. The proposed study is associated with XAI for Sign Language Recognition. The methodology uses an attention-based ensemble learning approach to create a prediction model more accurate. The proposed methodology used ResNet50 with the Self Attention model to design ensemble learning architecture. The proposed ensemble learning approach has achieved remarkable accuracy at 98.20%. In interpreting ensemble learning prediction, the author has proposed SignExplainer to explain the relevancy (in percentage) of predicted results. SignExplainer has illustrated excellent results, compared to other conventional Explainable AI models reported in state of the art.

INDEX TERMS Deep learning, computer vision, explainable AI, SignExplainer, classification, sign language, technological development.

I. INTRODUCTION

Arevolutionized era of Artificial Intelligence with Machine Learning and Deep Learning has demonstrated potential in a different sector. Over the one decade, Machine learning and deep Learning had a vast range of applications in research and industry, especially computer vision with deep learning has proven incredible results. Computer vision in fields like medicine, autonomous vehicles, agriculture, and remote sensing have little chance for failure [1]. Deep learning methods, computer vision, human-computer interface, and other related sub-fields have also illustrated compatible performance in various domains. Computer vision with deep learning has proven hard to fail for many tasks [2]. With

The associate editor coordinating the review of this manuscript and approving it for publication was Yongjie Li.

the availability of exclusive computing resources and a huge amount of learning dataset, deep learning can generate much more accurate results than before. With the good performance of machine learning and deep learning, artificial intelligence can achieve superhuman abilities. The world's social environment will transform dramatically due to artificial intelligence over the use of different platforms. These changes come with various ethical issues, which society will need to quickly adjust to influence the advances in a way that will lead to positive consequences. The complexity of deep learning models allows artificial intelligence to learn and react over complex data structures. Computer vision is one of the best approaches for image classification, segmentation, object detection, and many more applications [3].

Deep learning models prove excellent performances in sensitive areas like medical science, national defense, automation driving, finance, and many more, but these applications also need attention to trust-related problems. A system having promising results but with good interpretation is easier to trust [4]. The significant performance of computer vision task generates a huge number of parameters and links with the physical environment, which is extremely hard to explain. This complex learning structure generally considers as a "Black-Box" [5]. Since, the advancement of deep learning, especially computer vision in sensitive and critical sectors, the issue of transparency and interpretability is highly recommended. It's necessary to involve explainability in Artificial Intelligence generally referred to as Explainable Artificial Intelligence (XAI). A rapidly expanding field of study, XAI is quickly emerging as one of the more important components of Artificial Intelligence (AI) [6]. Research over XAI in the context of computer vision aims to extract or try to interpret the structure inside the black box. Additionally, it provides trust and interpretability to assist bias-free debugging over different computer vision applications like object detection, classification, and others. Interpretation from XAI models explains potential design flow or structures [7]. Figure 1 represents a functional comparison of AI and XAI, especially for reaction over predicted results by black box learning.

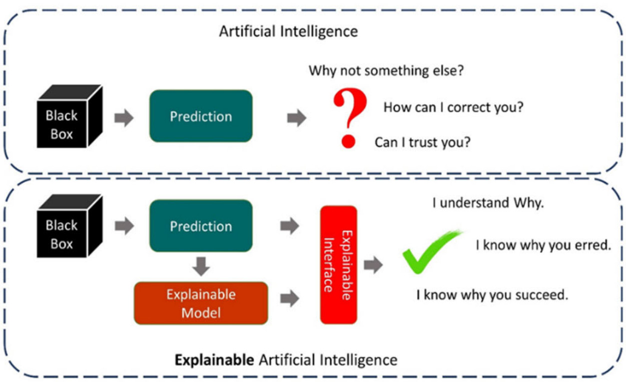


FIGURE 1. Architectural summary and analysis of artificial intelligence and explainable artificial intelligence.

For medical domain tasks like Sign language recognition, it is necessary to explain and relive the internal learning pattern. If the internal learning patent is correct, then it will increase trust in sign language recognition models. However, this explanation also provides misclassification error, leading to improvisation in the model or input scenario. Trust values are much more essential for sign language recognition to predict how the model will learn a given gesture-based sign [8]. The interpretability improves the methodology to predict the actual label. Because the generation of sign gestures may vary from person to person, in that case, there is a high possibility to recognize a different label. Sign language recognition with Explainable AI helps to improve the recognition model with various expectations, and also help the end user to understand the learning methodology of the deep learning model to recognize different sign gestures [9].

A sign language recognition system helps physically impaired people to communicate with the rest of the world. People having hearing impairment use gesture-based signs to express their emotions and thoughts. The majority of the contribution to generating a sign is a hand gesture, but to express proper meaning it will involve other non-manual body parts like the orientation of the head, the direction of eyes, eyebrows, and lips moment. XAI for sign language recognition helps to understand the predicted result, which may lead to improved accuracy of the model as well as users also get familiar with the generated ideal gesture of sign. Computer vision-based sign language recognition systems not only improve in terms of accuracy but also improve user trust [10].

This study proposed a threefold main contribution.

- First, Attention-based ensemble learning for sign language recognition.
- Second, the authors have introduced novel architecture using XAI for Sign language recognition.
- Finally, illustrate concrete evidence for interpretability and decision-driven approach of the proposed methodology with Explainable AI.

The rest of the article is designed as section II illustrates the recently published methodology for sign language recognition and XAI. Section III demonstrate the proposed methodology with deep learning and XAI. Section IV represents the simulation process and demonstrates the explainability and interpretability of the proposed architecture. Section V illustrates the evaluation and results discussions.

II. RELATED WORK

Kim et al. [11], introduce Concept Activation Vectors (CAVs), which translate a neural network's internal state into understandable ideas, which the author introduces. The important concept is to use a neural network's high-dimensional internal state as a tool rather than a hindrance. The authors have demonstrated the application of CAVs as a component of a method called Testing with CAVs (TCAV), which uses directional derivatives to gauge. How important a user-defined concept is to the categorization result, such as how much of a zebra prediction is influenced by the presence of stripes. We explain how CAVs may be used to evaluate predictions and generate knowledge for a standard image classification network and a medical application, putting concepts to the test in image categorization.

In this research [12], authors describe a unique technique that offers contrasting justifications for categorizing an input by a deep neural network or another black box classifier. Given an input, we find what needs to be simply and adequately present (viz. important object pixels in an image) to justify its classification and analogously, along with that minimally and necessarily absent (viz. certain background pixels) for the same. We contend that such explanations are typical in fields like criminology and health care because they are natural to people. A key aspect of an explanation that, to our knowledge, has not yet been formally identified by current explanation methods used to explain neural network predictions is minimally represented but critically not present. The authors have validated the proposed methodology over three datasets obtained from diverse domains; a brain activity strength dataset, a large procurement fraud dataset, and a handwritten digits dataset MNIST. In all three cases, we observe the effectiveness of our method in producing precise explanations that are also simple for specialists to comprehend and evaluate. [12].

Akula et al. [13], proposed the CoCoX model to explain the prediction generated by CNN classification. The author has proposed a fault-line model to identify minimum segmented-level features. Explanation from the CoCoX model was understandable to the technical and non-technical communities. The author has evaluated qualitative matrices like Justification Trust (JT), and Explanation Satisfaction (ES) to make performance understandable. The author has also compared the fault line model to other state-of-the-art models like LIME and LRP [13], author has successfully achieved 69.1 JT with CNN learning and Fault-Line Identification.

Contreras et al. [14], design Deep Explainer and Rule Extraction (DEXiRE), to make binary neural networks explainable. The proposed methodology uses rule extraction, which improves knowledge extraction from DL model (CNN) output. A final (global) rule set describing the general behavior of DL predictors can be created by integrating intermediate rule sets explaining the behavior of each concealed layer. They used BCWD, Banknote, and Prima diabetes datasets for the simulation of the proposed DEXiRE model. The number of words in the intermediate and final rule sets may be regulated precisely with DEXiRE. The rule Extraction model has achieved remarkable accuracy and fidelity 0.94 and 0.95 respectively in a very small amount of time (around 232 ms).

Patel et al. [15] water Potability prediction synthetic oversampling technique and Explainable AI. The author has used Synthetic Minority Oversampling Technique (SMOTE) method to classify water quality on the Kaggle dataset. The author has also compared the proposed architecture with other standard machine learning models like Design Tree, Gradient Boost, Support vector machine, Random Forest, and Ada Boost. The proposed methodology has achieved 81% remarkable accuracy. The author has also considered the lack of transparency issue for Machine Learning models. To determine the significance of the characteristics of the predicted result, Local Interpretable Model-agnostic Explanations (LIME) are used. The author has demonstrated the different available particles in water like Chloramines, Turbidity, Sulfate, and many more to justify results with Explainable AI, the proposed LIME model utilize to generate a result with the percentage of water particles.

Vermeire et al. [16] proposed a model-agnostic model ''Search for EviDence Counterfactual'' (SEDC) for image classification. The ''EdC'' explanation is an irreducible collection of characteristics that, if absent, would change the classification of the document. The SEDC additionally supports a single task for image explanation. The proposed methodology used image segmentation as a core component to interpret. The authors have the simulated model to compare different counterfactual classes and also compare with standard explainer models like SHAP and LIME. Simulation has used pre-train weights of MobileNet V2 to demonstrate the interpretation of the proposed SEDC model.

Goel et al. [17], a proposed technique to design ''counterfactual explanations''. Generally, it is used to justify by content area of the image, through the model that made the prediction. The methodology also encountered the problem of Minimum-Edit Counterfactual. A methodology work on input image trained by a computer vision model, to interpret the predicted class. The methodology used the MNIST dataset over the CNN model achieved 98.40% accuracy. The proposed training model has 2 convolutions and 2 FC (Fully connected) layers to generate a feature size of $4 \times 4 \times 40$. To generalize counterfactual explanations, the author has also experimented with Omniglot and Caltech-UCSD Birds dataset. Proposed technique working over Greedy Sequential Exhaustive Search model. The author has summarized the qualitative and quantitative results of the proposed technique.

Arras et al. [18], proposed a framework that provides, a controlled, selective, and realistic testbed for the prediction of deep neural networks. The proposed methodology uses the CLEVR-XAI dataset for simulation, there were around 140k questions in the CLEVR-XAI evaluation set. With 28 alternative solutions. The prediction issue is presented as a classification challenge. The author has used ten polling techniques to visualize the explanation evaluation over a round truth mask. The experiment section summarized the evaluation of different XAI methods like Guided Backprop, LRP, SmoothGrad, and other 7 methods [18]. The conclusive study finds that LRP performed much better compared to another method over the proposed (CLEVR-XAI) benchmark dataset. Table 1 represent comparative analysis over different explainable model to predict result by black-box learning, analysis also represents a statistical comparison to justify trust and confidence.

TABLE 1. Comparative analysis of state-of-the-art Explainable AI model overconfidence and justified trust value.

Author	Model	Justified Trust	Confidence
Zhou et al. 2016 [19]	CAM	37.1% ± 3.9%	3.2 ± 1.8
Selvaraju et al. 2017 [20]	Grad-CAM	39.1% ± 2.1%	3.7 ± 1.2
Ribeiro, Singh, and Guestrin 2016 [21]	LIME	42.1% ± 3.1%	3.1 ± 2.2
Kim et al. 2018 [11]	TCAV	55.1% ± 3.3%	3.9 ± 2.8
Dhurandhar et al. 2018 [12]	CEM	61.1% ± 2.2%	4.8 ± 1.6
Goyal et al. 2019 [17]	CVE	64.5% ± 3.7%	4.1 ± 2.3
Akula et al. 2020 [13]	CoCoX	70.5% ± 1.3%	5.7 ± 1.1
Vermeire et al. 2022 [16]	SEDC	71.4% ± 2.1%	6.1± 1.0

III. MATERIALS AND METHODS

The proposed architecture used an Explainable AI-based methodology for sign language recognition with DeepExplainer. Which use to predict and validate generated output with learning interpretability. The proposed methodology uses SHAP (Shapley Additive exPlanations) [18] to interpret framework prediction. A global interpreter SHAP is used over LIME [22], to interpret the effect of the single feature on the target variable. SHAP framework utilizes various explainability methods for better interpretation of model prediction. The proposed methodology is divided into three major stages i) Ensemble learning, ii) Prediction of learning, iii) Sign Explainer, and interpret the results. Figure 2 shows the sequential flow of the proposed model.

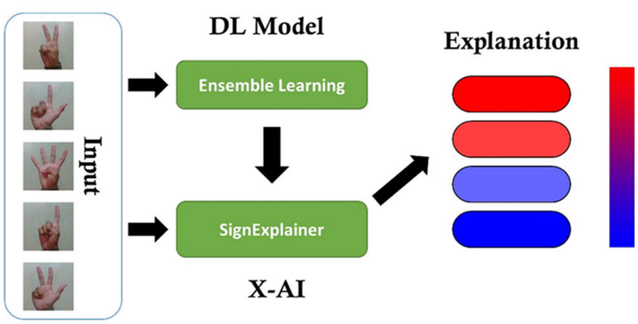


FIGURE 2. Sequential process architecture of proposed methodology.

A. ENSEMBLE LEARNING

Every custom Deep Learning model is based on training-based learning and must necessarily stage to learn deep features. Especially, when the task was related to computer vision, proper model training is necessary. The proposed methodology used ensemble learning with an attention model. Figure 3 represents an ensemble attention-based model for sign language recognition. The proposed methodology uses a bagging-based ensemble model to learn the associated feature of sign images. Attention-based Ensemble learning mainly divides into two categories, multi-head ensemble and attention-based ensemble [23]. Figure 3 demonstrate the different way of attention-based ensemble learning. Algorithm 1 represents the architectural structure of the proposed ensemble learning approach with the bagging concept.

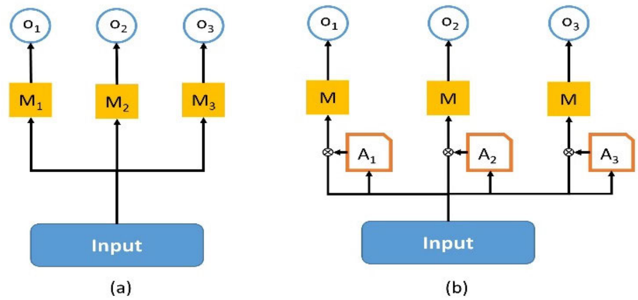


FIGURE 3. A different way to use attention for ensemble learning, (a) represents a multi-head ensemble with different feature embedding parameters, (b) represents the same feature embedding with different feature learning.

The proposed methodology used ensemble learning. Which is mainly divided into two parts. The first one is ResNet50 with a 23.521M parameter as part of the convolution learning module. ResNet50 is used to reduce the vanishing gradient problem. Generally, in a deep convolution network loss function is shrunk to zero after several iterations. With the help of the ResNet network, gradients can be directed to skip connections from previous layers to the next filter layer. The linear learning of residual network can be considered as equation 1. [24], where $G(x, and \{W_i\})$ stand for mapping of residual learning, while W_s and x stand for projection square matrix of x dimension.

$$\eta = G(x, \{W_i\}) + W_s + x \tag{1}$$

Another component of ensemble learning is the attention module, which can be designed with two associated modules as feature extraction module F(x) and attention module A(x). The feature extraction module was designed with a pro-layer perceptron model, and generalized as equation 2 [23]. And the attention weights were calculated as equations 3 and 4, where h_e and h_d stand for encoder and decoder weights.

$$F(x) = h_i(h_{ii-1}(\ldots(h_2(h_1(x))) \tag{2}$$

$$\gamma = \tanh(W * h_e + W * h_d) \tag{3}$$

$$A(x) = Softmax(\gamma) \tag{4}$$

The global feature embedding model $G(x)$ (equation 5), for the embedding module. Authors have proposed

Algorithm 1 Pseudo-Code for Proposed Ensemble Learning Architecture (Bagging based)

Input: Training Image set I
Output: Interpretation_Index
$K \leftarrow Conv\_Layer(ResNet(i))$
$l \leftarrow Class\_Labels\ \{0,1,2\ldots,A,B,\ldots Z\}$
$G \leftarrow Ensemble\_feature(l)$
$C \leftarrow Num\_Classes(l)$
for $k \in \{1, \ldots, K\}$ **do**
  for $C \in \{0, \ldots, C\}$ **do**
    $D_c = Conv(I_{c0} * I_{c1} * \ldots.. * I_{cn})$
    $f_c = D_0 \cup D_1 \cup \ldots\ldots. \cup D_n$
  end for
$G(k) = MLP(f_c)$
end for
$G(x) = softmax((G1(x) + G2(x) + \ldots + Gk(x)) / K)$
#Feature Explainer:
procedure Sign_Ex(g(x), l)
  $i \leftarrow max\_val(int)$
  $Create \Pi\ \ for\ collections$
  for i ∈ g(x) **do**
    for each $\pi \in \{\pi_0 \ldots\ldots \pi_1\}$ **do**
      $Calculate \pi_i$;
      $\pi_O \leftarrow \Delta(\pi_i)$
    end for
  $\Upsilon \leftarrow$ evaluate(π_O, l)
  end for
return(index ← max_val(Υ))
end procedure

three-dimension blob channel to recognize input images in an RGB channel. The attention feature and convolution feature are associated with the final feature vector generation and it was forwarded to a fully connected DCNN network for classification. Figure 4 represents the conceptual architecture representation of the proposed ensemble learning with the attention model.

$$G(x) = \sum F(x) \otimes A(x) \tag{5}$$

B. CLASSIFICATION AND PREDICTION

The output from the fully connected layer is further processes for classification and prediction. The authors have implemented multi-layer perceptron (MLP) [25] to classify sign language. The proposed methodology uses DFFN (Deep Forward Neural Network) to recognize gesture signs from input images. ReLU activation was implemented in the final layer of the deep network for sign recognition, and it can be calculated as equation (6), where (W_1, W_2) are different weights and (b_1, b_2) as bias.

$$DFNN = ReLU\ (W_{1x} + b_1)\ W_2 + b_2 \tag{6}$$

The authors have utilized NumPy and Scikit-learn [26] for evaluation and visualization. The class-wise performance score has been calculated, and accuracy, precision, recall, and F1-Score were calculated to analyze ensemble model performance. Performance standards have been calculated as per equations 7 to 10. [27], [28].

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{7}$$

$$Precision = TP\big/(TP + FP) \tag{8}$$

$$Recall = TP\big/(TP + FN) \tag{9}$$

$$F1 - Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{10}$$

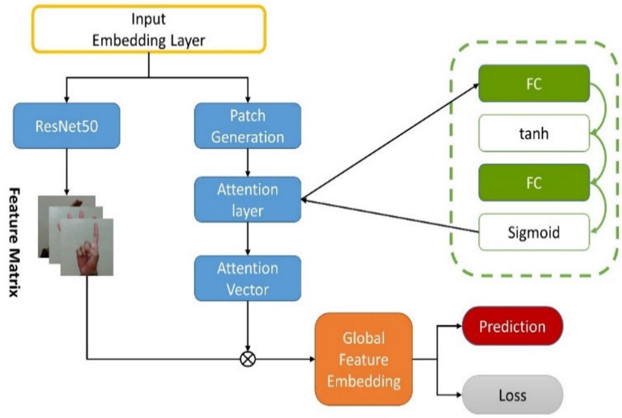


FIGURE 4. Proposed ensemble learning architecture with ResNet50 and attention model, to learn embedded features with global feature embedding method, (FC stands for fully connected layer).

C. SIGNEXPLAINER

Interpretation and explainable techniques involved with black-box deep learning models fall under two categories, model specific or agnostic. This section focuses on the design of SignExplainer an agnostic interpretability technique, that can be applied to any black-box deep-learning model to interpret gesture-based signs. SHAP [29] is among the most utilized interpretability methods for deep learning-based methods. SHAP can construct interpretations for multi-class classifier responses. SignExplainer uses Sign-specific Xconcept to generate a fault line explanation. Let's assume that δ_{pred} and δ_{alt} can be Xconcept for E_{alt} and E_{alt} respectively where E stands for the actual class. Based on Xconcept, line prediction can be calculated as equation 11 [30].

$$\Psi(E_{pred}, E_{alt}) \leftarrow \min_{\delta_{pred}, \delta_{alt}} \alpha\left(\delta_{pred}, \delta_{alt}\right) + \beta\left||\delta_{pred}|\right| + \lambda\,||\delta_{alt}|| \tag{11}$$

The proposed Methodology designs DeepExplainer as an additive feature attribution method with accuracy and missingness. DeepExplainer combines the SHAP value computed for a smaller component of the ensemble network and calculates it as equation 12, [31]. Where, $o = f(x) - f(r)$ and $x_i = x_i - x_r$, $r$is the reference input, while $f(x)$ is the model output,

$$O = \sum\nolimits_{i=1}^{n} Cx_i * \Delta o \tag{12}$$

IV. EXPERIMENTS AND RESULT

A. DATASET

The authors have evaluated SignExplainer with ensemble learning on Indian Sign Language Dataset [32]. The dataset used for simulation consists of 36 Indian Sign classes having digits (0-9) and an alphabet (A-Z). The dataset consists of approximately 1200 images per class, with 3 channel images. Along with Indian Sign Language (ISL) dataset, the authors have also experimented with other static datasets like American Sign Language (ASL) [33], and Bangla Sign Language (BSL) [34]. Property of datasets described in Table 2.

TABLE 2. Statistical representation of different sign language datasets used in the simulation.

Dataset	Avg. Resolution	Classis	Avg. Image per class
Indian Sign Language (ISL) [32]	250 x 250	36	1200
American Sign Language (ASL) [33]	400 x 400	35	840
Bangla Sign Language (BSL) [34]	171 x 166	33	654

B. DATA AUGMENTATION

The proposed simulation uses data augmentation to make the model more generalized for feature learning. Data augmentation is also used to balance training image samples and improve robustness for learning variability over the different images, making the model more generalized toward real-time scenarios. Direct image inference may yield biased findings due to particular transformations and noise associated with equipment and surroundings. Image augmentation must be used to achieve more reliable and robust prediction to improve accuracy and prevent overfitting. The authors have implemented i) Geometric transformations as random horizontal flip, random rotation with +0.2 to -0.2, and zooming by 1.5% to 2.5%. ii) Color space transformations as random RGB change and Brightness by 0.5%. Figure 5 represents the sample of the augmented training dataset.

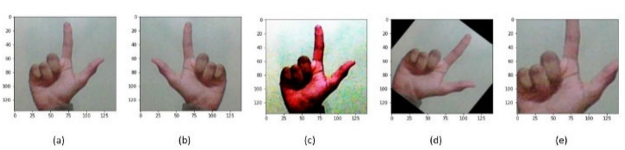


FIGURE 5. Input Sign image augmentation, (a) original image, (b) horizontal flip, (c) color transformation, (d) random rotation, (e) zooming.

C. SIMULATION DETAILS

The authors have implemented training of an ensemble learning module on the ISL dataset [32]. TensoFlow-Keras has been used for the design of the proposed methodology. The proposed ensemble methodology has achieved 98.20 % accuracy with extracted features from attention and the ResNet50 model. Model training was divided with 0.2 train-test split ratios (80:20) for all experiments, with an image size of (72, 72, 3) and a batch size of 16. The model was simulated with 0.3 as a dropout ratio and a 0.001 learning rate with the Adam optimizer. Table 3 demonstrate superior performance over other standard Convolution networks, additionally, the best performance was observed by the proposed Attention-based ensemble model. The proposed methodology has achieved significant accuracy over 50 learning epochs, as shown in Figure 6.

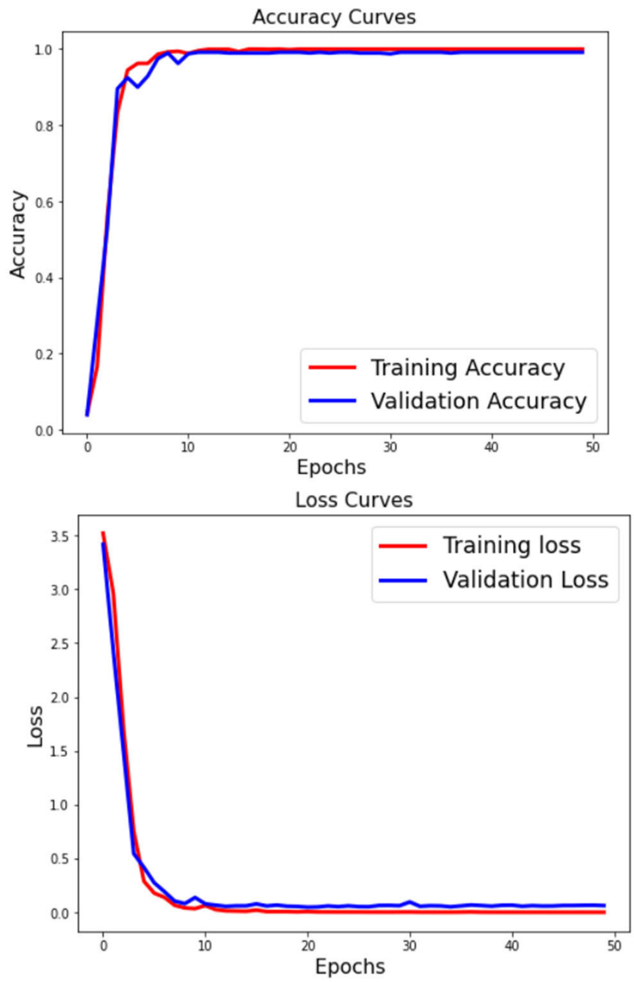


FIGURE 6. Accuracy and loss curve for Indian Sign Language recognition using Attention-based Ensemble learning.

D. INTERPRETATION WITH SIGNEXPLAINER

The proposed methodology simulates SignExplainer to generate a model prediction and explain the correctness of the prediction. The simulation uses OpenCV for masking the input images and passes them to the "blur (128,128)" method, which is responsible to mask the predicted image output with inpaint-telea value. The authors have created

TABLE 3. Performance analysis with state-of-the-art models for Image classificatio.

Model	Accuracy	Precision	Recall	F1-Score
CNN [35]	92.60%	0.92	0.92	0.91
VGG16 [36]	97.55%	0.98	0.97	0.97
EfficientNet V2 [37]	96.42%	0.96	0.96	0.95
Ensemble (ResNet50 + Attention)	98.20%	0.98	0.98	0.97

SignExplainer with adaptive feature abstraction, which compares with and without x-features. X- Features are the associative contribution of ensemble learning features. Prediction function of SignExplainer, which is working as a masked feature. The authors have passed sign images with the Explainer object to generate SHAP values, and Figure 6 represents the plot.

The interpretation plot has been taken with 4 flips over 1,000 evaluations as (max_eval=1000) for the Explainer object (shown in figure 7). The gradient bar prediction represents the prediction's relevance interpretation, red stands for the maximum, and blue stands for the minimum. Table 4 represents the performance of other basic XAI models to interpret the ensemble model prediction output over the Indian Sign Language dataset.

TABLE 4. Statistical performance comparison of different models for Interpretation over ISL dataset (where TRP is True Positive Rate, FNR is False Negative Rate, PPV is Positive Predictive value, and FDR is False Discovery Rate).

Explainer Model	TRP	FNR	PPV	FDR
DeepLine [38]	80.8	19.2	70.8	29.2
Lime [21]	82.4	17.6	72.9	27.1
DeepLIFT [39]	79.1	20.9	74.8	25.2
SignExplainer	87.3	12.7	78.5	21.5

We have demonstrated the remarkable result of explanation over sign language, especially for Indian signs. To ensure the robustness of the proposed SignExplainer with an ensemble learning model, the author has evaluated the proposed methodology over other static and standard sign language datasets like American Sign Language ASL) and Bangla Sign Language (BSL), statistical comparison describe in Table 5.

The prediction score of SignExplainer for the test sign image is demonstrated in Figure 8. SignExplainer helps to understand and recognizes why the model recognizes the data instance as it has. The first image is from the testing dataset as a significant gesture of "4." The top of all predictions shows the matching value. Red dots represent high relevance while Blue dots represent low relevance. Based on the high

TABLE 5. Performance analysis of SignExplainer over different static Sign Language Datasets.

Dataset	Accuracy with Ensemble learning	Justified Trust	Confidence
ISL	98.2	89.1 ± 2.2	7.3 ± 1.4
ASL	98.0	87.4 ± 1.9	6.7 ± 2.0
BSL	96.7	73.6 ± 2.7	6.4 ± 1.8

relevance of feature attribution it's easy to interpret how the model was learned to predict sign language. The presence of a red pixel over the corresponding area of the hand gesture increases the prediction probabilities.

FIGURE 7. Support feature for SignExplainer over Indian Sign Language Recognition, (a few samples have been taken to maintain article readability).

V. RESULTS ANALYSIS AND DISCUSSION

A Computer vision-based model to learn and interpret the prediction was proposed by this study. The authors have proposed a sequential (two-phase) methodology from learning from the ensemble model to interpretation of the predicted result, with the SignExplainer model. The authors have also implemented the proposed architecture for Indian Sign Language (ISL). Experiment also extends to other static sign languages like American Sign Language (ASL), and

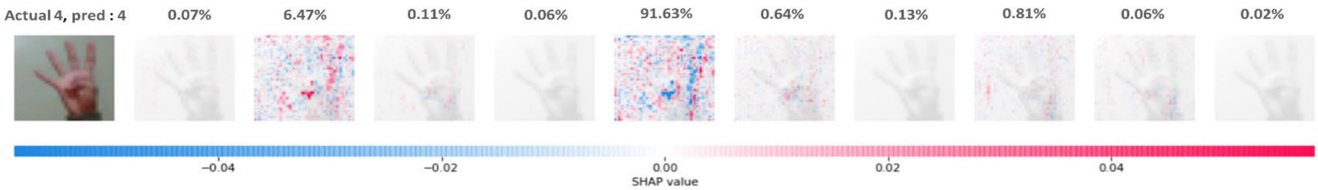


FIGURE 8. Representation of SignExplainer to interpret sign gesture with prediction value and class (class stars from 0-9 in left to right).

Bangla Sign Language. This study proposed and demonstrated attention-based ensemble learning with ResNet50 and Self-attention model. The proposed architecture was able to achieve 98.20 percent remarkable accuracy for ISL, and also compare with other computer vision state-of-the-art models. The second phase of the study demonstrated the interpretation of the learning model. The authors have used the SignExplainer model to extract masked values from the black-box model.

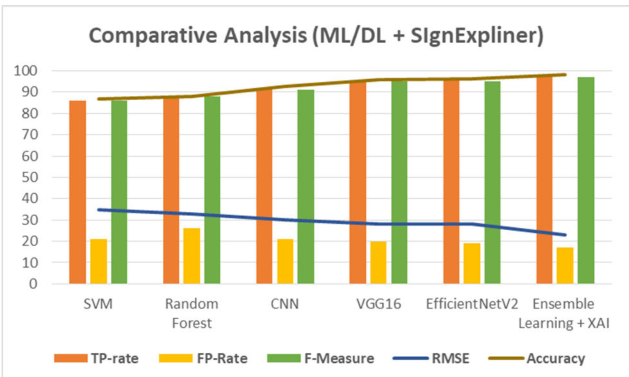


FIGURE 9. Comparative analysis of proposed methodology with other deep learning State-of-the-art methodology.

The proposed SignExplainer uses fault line calculation to interpret the correctness of the predicted sign image. The result section also demonstrates the achieved result by SignExplainer, and also compare it with other conventional XAI model. The author has also evaluated TP-rate and FP-rate for the proposed model, and it's found remarkable with other black box deep learning models as 0.98 and 0.17 respectively. Figure 9 represents a comparative analysis of the proposed architecture (ensemble learning + SignExplainer) with other deep learning models like SVM [40], Random Forest [41], CNN [35], VGG16 [36], and EfficientNetV2 [37]. The evaluation matrix was calculated with a True-False positive rate, F-measures, and RMSE (Root Mean Square Error) value. The statistical analysis represents the proposed associative architecture is more accurate than other standard machine learning and deep learning models (shown in Figure 9). The authors have also analyzed other deep learning object detection models like R-CNN [42], Faster R-CNN [43], and Single Shot Detector (SSD) [44] with VGG16 [45] as the backbone over the proposed Attention-based Ensemble model. A comparative analysis of deep learning detection models was illustrated

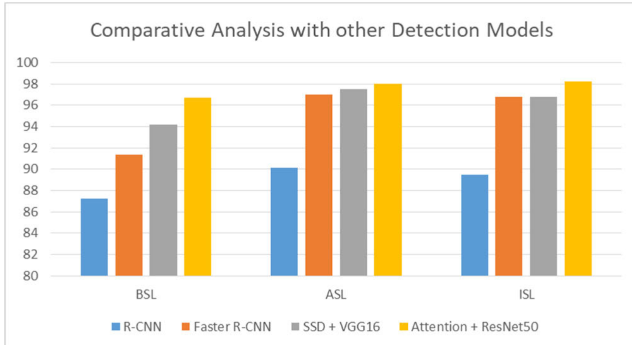


FIGURE 10. Comparative accuracy analysis of proposed methodology with other deep learning State-of-the-art object detection models.

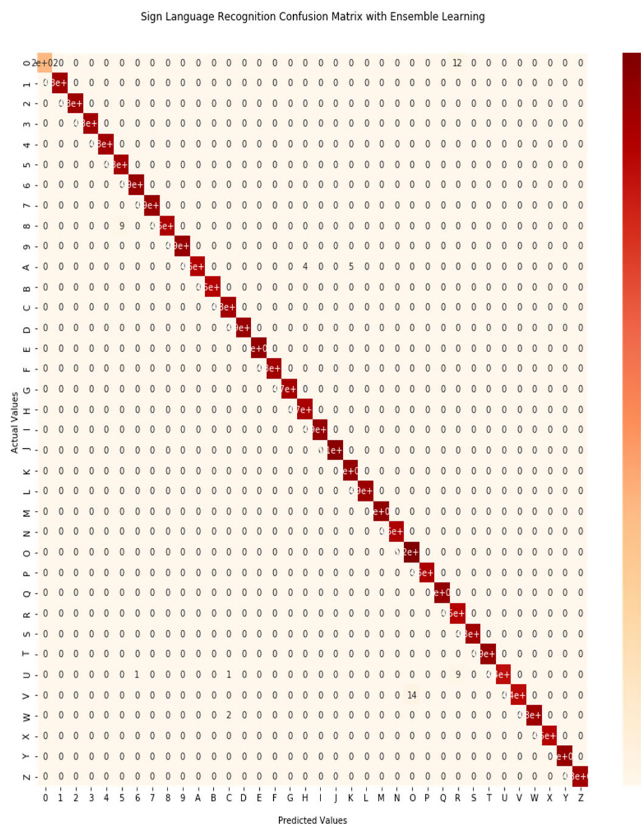


FIGURE 11. Confusion Matrix for Static Indian Sign Language using Ensemble Learning with ResNer50.

in Figure 10. Figure 11 illustrates the confusion matrix of the proposed ensemble learning methodology for the static Indian Sign Language dataset.

VI. CONCLUSION

The era of Explainable AI growing exponentially, to overcome trust and transparency issues of deep learning models. Especially tasks relevant to Computer vision or NLP must require interpreting predicted results over critical sectors. The review has explored different XAl methodologies like LRP, LIME, SHAP, and SmoothGrad over relevant computer vision applications. This study has proposed Sign Language Recognition to make explainable artificial intelligence. Ensemble learning-based architecture was proposed to recognize sign gestures from sign images. Ensemble weights were passed to the proposed SignExplainer to generate statistical values like TP-rate and FP-rate, to evaluate the correctness of the proposed SignExplainer. This study also evaluated ensemble learning with another deep learning model for image classification. The proposed study also evaluates the performance of SignExplainer over other benchmark static sign language datasets like ASL and BSL, and it also achieves remarkable performance. The proposed study also simulates additional machine learning and deep learning models like Decision tree, Random Forest, VGG16, and EfficientNetV2, and evaluates the performance of SignExplainer. Ensemble learning and other deep learning models were also performed well over SignExplainer to interpret predicted signs with proper statistical values. The proposed work can be extended to other static Sign Languages as well as isolated Sign Languages. The proposed methodology can be enhanced for real-time or portable Sign Language Recognition with acceptable interpretations.

ACKNOWLEDGMENT

This research was funded by Princess Nourah bint Abdulrahman University and Researchers Supporting Project number (PNURSP2023R346), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors would also like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

REFERENCES

[1] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: An analytical review," *WIREs Data Mining Knowl. Discovery*, vol. 11, no. 5, p. e1424, 2021.

[2] Y. Yuan and Y. Lo, "Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 519–526, Mar. 2019, doi: 10.1109/jbhi.2017.2787487.

[3] A. Gramegna and P. Giudici, "SHAP and LIME: An evaluation of discriminative power in credit risk," *Frontiers Artif. Intell.*, vol. 4, Sep. 2021, Art. no. 752558.

[4] F. Afza, M. A. Khan, M. Sharif, S. Kadry, G. Manogaran, T. Saba, I. Ashraf, and R. Damaševičius, "A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection," *Image Vis. Comput.*, vol. 106, Feb. 2021, Art. no. 104090.

[5] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/e23010018.

[6] K. V. Dudekula, H. Syed, M. I. M. Basha, S. I. Swamykan, P. P. Kasaraneni, Y. V. P. Kumar, A. Flah, and A. T. Azar, "Convolutional neural network-based personalized program recommendation system for smart television users," *Sustainability*, vol. 15, no. 3, p. 2206, Jan. 2023.

[7] M. Baldeon Calisto and S. K. Lai-Yuen, "AdaEn-net: An ensemble of adaptive 2D–3D fully convolutional networks for medical image segmentation," *Neural Netw.*, vol. 126, pp. 76–94, Jun. 2020, doi: 10.1016/j.neunet.2020.03.007.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[9] J. Ganesan, A. T. Azar, S. Alsenan, N. A. Kamal, B. Qureshi, and A. E. Hassanien, "Deep learning reader for visually impaired," *Electronics*, vol. 11, no. 20, p. 3335, Oct. 2022.

[10] D. Kothadiya, C. Bhatt, K. Sapariya, K. Patel, A.-B. Gil-González, and J. M. Corchado, "Deepsign: Sign language detection and recognition using deep learning," *Electronics*, vol. 11, no. 11, p. 1780, Jun. 2022, doi: 10.3390/electronics11111780.

[11] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. Int. Conf. Mach. Learn.*, Mar. 2023, pp. 2668–2677. [Online]. Available: http://proceedings.mlr.press/v80/kim18d.html

[12] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," 2018, *arXiv:1802.07623*.

[13] A. Akula, S. Wang, and S.-C. Zhu, "CoCoX: Generating conceptual and counterfactual explanations via fault-lines," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 3, pp. 2594–2601, doi: 10.1609/aaai.v34i03.5643.

[14] V. Contreras, N. Marini, L. Fanda, G. Manzo, Y. Mualla, J.-P. Calbimonte, M. Schumacher, and D. Calvaresi, "A DEXiRE for extracting propositional rules from neural networks via binarization," *Electronics*, vol. 11, no. 24, p. 4171, Dec. 2022, doi: 10.3390/electronics11244171.

[15] J. Patel, C. Amipara, T. A. Ahanger, K. Ladhva, R. K. Gupta, H. O. Alsaab, Y. S. Althobaiti, and R. Ratna, "A machine learning-based water potability prediction model by using synthetic minority oversampling technique and explainable AI," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–15, Sep. 2022, doi: 10.1155/2022/9283293.

[16] T. Vermeire, D. Brughmans, S. Goethals, R. M. B. de Oliveira, and D. Martens, "Explainable image classification with evidence counterfactual," *Pattern Anal. Appl.*, vol. 25, no. 2, pp. 315–335, Jan. 2022, doi: 10.1007/s10044-021-01055-y.

[17] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *Proc. 36th Int. Conf. Mach. Learn.*, May 2019, pp. 2376–2384, Accessed: Mar. 2023. [Online]. Available: https://proceedings.mlr.press/v97/goyal19a.html

[18] L. Arras, A. Osman, and W. Samek, "CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations," *Inf. Fusion*, vol. 81, pp. 14–40, May 2022, doi: 10.1016/j.inffus.2021.11.008.

[19] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929, Accessed: Mar. 6, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Zhou_Learning_Deep_Features_CVPR_2016_paper.html

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier" in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[22] X. Shen, K. Lu, S. Mehta, J. Zhang, W. Liu, J. Fan, and Z. Zha, "MKEL: Multiple kernel ensemble learning via unified ensemble loss for image classification," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 4, pp. 1–21, Aug. 2021.

[23] W. Kim, B. Goyal, K. Chawla, J. Lee, and K. Kwon, "Attention-based ensemble for deep metric learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV*, 2018, pp. 736–751.

[24] B. Chen and W. Deng, "Deep embedding learning with adaptive large margin N-pair loss for image retrieval and clustering," *Pattern Recognit.*, vol. 93, pp. 353–364, Sep. 2019, doi: 10.1016/j.patcog.2019.05.011.

[25] D. R. Kothadiya, C. M. Bhatt, T. Saba, A. Rehman, and S. A. Bahaj, "SIGNFORMER: DeepVision transformer for sign language recognition," *IEEE Access*, vol. 11, pp. 4730–4739, 2023, doi: 10.1109/access.2022.3231130.

[26] J. Mueller and L. Massaron, *Python for Data Science*. Hoboken, NJ, USA: Wiley, 2019.

[27] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using real-sense," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP*, Jul. 2015, pp. 166–170.

[28] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2015, pp. 572–578.

[29] S. Knapič, A. Malhi, R. Saluja, and K. Främling, "Explainable artificial intelligence for human decision support system in the medical domain," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 3, pp. 740–770, Sep. 2021, doi: 10.3390/make3030037.

[30] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, "Evaluating XAI: A comparison of rule-based and example-based explanations," *Artif. Intell.*, vol. 291, Feb. 2021, Art. no. 103404.

[31] F. Gabbay, S. Bar-Lev, O. Montano, and N. Hadad, "A LIME-based explainable machine learning model for predicting the severity level of COVID-19 diagnosed patients," *Appl. Sci.*, vol. 11, no. 21, p. 10417, Nov. 2021.

[32] D. R. Kothadiya. (Oct. 2022). *Deepkothadiya/STATIC_ISL: Static Indian Sign Language Dataset Having Sign of Digit and Alphabet*. [Online]. Available: https://github.com/DeepKothadiya/Static_ISL

[33] Thakur. (May 2019). *American Sign Language Dataset*. [Online]. Available: https://www.kaggle.com/datasets/ayuraj/american-sign-language-dataset

[34] S. M. Rayeed. (Aug. 2021). *Bangla Sign Language Dataset*. [Online]. Available: https://www.kaggle.com/datasets/rayeed045/bangla-sign-language-dataset

[35] T. Saba, M. A. Khan, A. Rehman, and S. L. Marie-Sainte, "Region extraction and classification of skin cancer: A heterogeneous framework of deep CNN features fusion and reduction," *J. Med. Syst.*, vol. 43, no. 9, Jul. 2019, doi: 10.1007/s10916-019-1413-3.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[37] B. Li, B. Liu, S. Li, and H. Liu, "An improved EfficientNet for Rice germ integrity classification and recognition," *Agriculture*, vol. 12, no. 6, p. 863, Jun. 2022, doi: 10.3390/agriculture12060863.

[38] Y. Heffetz, R. Vainshtein, G. Katz, and L. Rokach, "DeepLine: AutoML tool for pipelines generation using deep reinforcement learning and hierarchical actions filtering," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 2103–2113.

[39] H. Chen, S. Lundberg, and S.-I. Lee, "Explaining models by propagating Shapley values of local components," in *Explainable AI in Healthcare and Medicine*. Cham, Switzerland: Springer, 2020, pp. 261–270. [Online]. Available: https://link.springer.com/book/10.1007/978-3-030-53352-6?page=2#toc, doi: 10.1007/978-3-030-53352-6.

[40] A. Razaque, M. Ben Haj Frej, M. Almi'ani, M. Alotaibi, and B. Alotaibi, "Improved support vector machine enabled radial basis function and linear variants for remote sensing image classification," *Sensors*, vol. 21, no. 13, p. 4431, Jun. 2021, doi: 10.3390/s21134431.

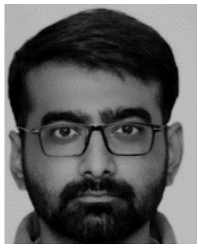
[41] Z. Noshad, N. Javaid, T. Saba, Z. Wadud, M. Saleem, M. Alzahrani, and O. Sheta, "Fault detection in wireless sensor networks through the random forest classifier," *Sensors*, vol. 19, no. 7, p. 1568, Apr. 2019.

[42] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," 2021, *arXiv:2108.05699*.

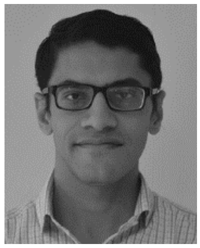
[43] Y. Liu, "An improved faster R-CNN for object detection," in *Proc. 11th Int. Symp. Comput. Intell. Design (ISCID)*, vol. 2, Dec. 2018, pp. 119–123.

[44] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," *Proc. Comput. Vis. (ECCV)*, 2016, pp. 21–37.

[45] A. T. Azar, Z. I. Khan, S. U. Amin, and K. M. Fouad, "Hybrid global optimization algorithm for feature selection," *Comput., Mater. Continua*, vol. 74, no. 1, pp. 2021–2037, 2023.



DEEP R. KOTHADIYA received the bachelor's and master's degrees in computer science and engineering from Gujarat Technological University. He is currently pursuing the Ph.D. degree with the Charotar University of Science and Technology (CHARUSAT). He is also an Assistant Professor with the U & P U Patel Department of Computer Engineering, Chandubhai S. Patel Institute of Technology, CHARUSAT. He is also a Research Scholar with CHARUSAT, and Prince Sultan University, Riyadh, Saudi Arabia. He has already published many research papers, including one SCI-indexed paper. He is also a Technical Reviewer of *International Journal of Computing and Digital Systems*.



CHINTAN M. BHATT (Member, IEEE) was an Assistant Professor with the CE Department, CSPIT, CHARUSAT, for 11 years. He is currently an Assistant Professor with the Department of Computer Science and Engineering (CSE), School of Technology, Pandit Deendayal Energy University (PDEU). He is the author or coauthor of more than 80 publications in the areas of computer vision, the Internet of Things, and fog computing. He was involved in successful organization of few special issues in SCI/Scopus journals. He has won several awards, including the CSI Award and the Best Paper Award for his CSI articles and conference publications.

AMJAD REHMAN (Senior Member, IEEE) received the Ph.D. and postdoctoral degrees (Hons.) from the Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Malaysia, with a specialization in forensic documents analysis and security, in 2010 and 2011, respectively. He is currently a Senior Researcher with the Artificial Intelligence and Data Analytics Laboratory, Prince Sultan University, Riyadh, Saudi Arabia. He is also a PI in several funded projects and also completed projects funded from MOHE Malaysia, Saud Arabia. His research interests include data mining, health informatics, and pattern recognition.

FATEN S. ALAMRI received the Ph.D. degree in system modeling and analysis in statistics from Virginia Commonwealth University, USA, in 2020. Her Ph.D. research was in Bayesian dose response modeling, experimental design, and nonparametric modeling. She is currently an Assistant Professor with the Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdul Rahman University. Her research interests include spatial area, environmental statistics, and brain imaging.

TANZILA SABA (Senior Member, IEEE) received the Ph.D. degree in document information security and management from the Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia, in 2012. She is currently the Associate Chair of Information Systems Department, College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia. Her primary research interests include medical imaging, pattern recognition, data mining, MRI analysis, and soft-computing.