OSTEOARTHRITIS DISEASE DETECTION USING DEEP LEARNING ARCHITECTURES

[1] Dr. Beebi Naseeba, [2] Gudiguntla Vyshnavi, [3] Thanneeru Priyanka
[2,3] Student B-Tech CSE-AI, [1]Assistant Professor,
School of Computer Science and Engineering (SCOPE)
VIT-AP University, India
beebi.naseeba@vitap.ac.in, gvyshnavi693@gmail.com, priyankat3469@gmail.com

***Abstract*—The disease osteoarthritis (OA) is the most prevalent type of condition affecting the musculoskeletal system and its main cause of elderly people's decreased mobility. Evaluation of visual subjectivity still plagues OA. [1] An accurate prediction is a necessary first step to effectively identify and prevent severe osteoarthritis. Experts frequently use a manual examination of patient medical photos, typically gathered in hospitals, to diagnose osteoarthritis. For patients, determining whether osteoarthritis is present takes some time. The present studies also concentrate on using image-based deep learning systems to automatically identify osteoarthritis. Medical photographs of the patients are required, which necessitates hospital visits from the patients. In this study, a deep neural network was employed to identify the presence of osteoarthritis. Our study results on data from the publicly available Kaggle that the established method may provide results for the early identification of OA in the knee.[2]**

Keywords— L2 regularization, EfficientNetV2L, MobileNet, Batch Normalization, Osteoarthritis, Dropouts, Learning Rate, Dense Layers

I. INTRODUCTION

Joint failure is a widespread, age-related, chronic, and slowly progressing joint condition called osteoarthritis. The joints that are most usually affected by osteoarthritis are those that are overused throughout life, such as the knees, neck, low back, and little joints at the ends of your fingers. The cartilage, a cushion made of silky rubber that covers the ends of the bones in the joint, is affected by this disorder. It has a few cells that secrete a complex matrix made of proteoglycans and hyaluronic acid.

Throughout life, worn-out matrix elements are constantly replaced even though cartilage has a very limited capacity for repair and adaptation. Menisci and other joint components are damaged, as well as the remodeling of 2 bones when cartilage breaks down in OA. OA is currently thought to be a disease brought on by the joints' dynamic response to various biomechanical and biochemical variables. When cartilage deterioration outpaces healing, it happens. The ligaments around the joint loosen, the joint capsule thickens, the joint fluid thins out, and the muscle deteriorates and weakens. Joint failure, a disorder similar to heart failure or kidney failure, is the result of these alterations. There is no discomfort in the cartilage. Bone and other tissues within and around the joint cause pain in OA. In OA, crystals and joint edoema are 2 major sources of discomfort.[3]

Around 10% of men and 13% of women over 60 have symptomatic knee OA, a disease that worsens with age and increases in frequency and prevalence. The number of persons who have OA that is the most severe in their age group is growing as the population ages. A primary contributor to morbidity and disability, OA also has high socioeconomic implications. The cost of arthritis in the United States was projected to be $336 billion in 2004, accounting for 3% of the country's GDP, with osteoarthritis (OA) being the most prevalent type. About 4-6% of adults in India have osteoarthritis, which is one of the top 5 chronic disorders.

According to statistics, India annually reports more than 10 million instances of osteoarthritis. By 2025, the nation is anticipated to become the world's osteoarthritis capital with more than 60 million cases.

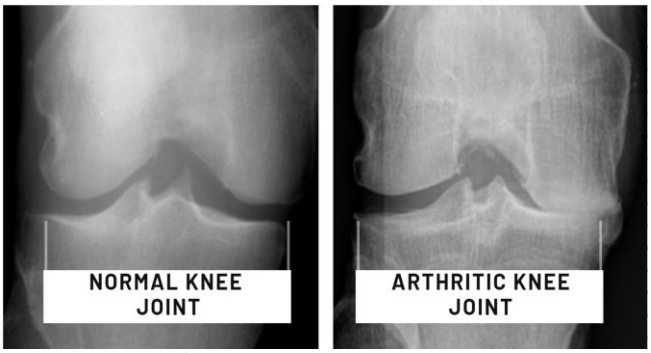


Fig. 1. *Knee Joints of a normal and diseased patient*

The most prevalent musculoskeletal disorder in the world is OA. According to the literature, hip and knee OA rank as the eleventh-highest worldwide disability factor1, placing a significant economic burden on society. According to reports, the average annual cost per patient for OA treatments is 19,000 euros. The current clinical failure to consistently detect the disease at an early stage, when it might still be possible to stop its course or at least lessen the severity of its future handicap, accounts for a portion of these expenditures. The only choices left to extend the patients' healthy years of life are an early diagnosis and behavioral interventions3 because there is no effective treatment for OA other than complete joint replacement surgery at the advanced stage.

Clinically, OA can be diagnosed early, but this needs the costly magnetic resonance imaging (MRI) that is now only available in specialized facilities or private practice. Additionally, this modality misses changes in bone architecture, which could be a sign of early OA progression.

II. LITERATURE SURVEY

S.No	Author	Model	Approach	Accuracy
1.	A. Chetouani, E. Lespessailles, Y. Nasser, R. Jennane, M. and E. Hassouni	discriminative regularised auto-encoder (DRAE)	A system that is based on auto-encoders	0.8173
2.	Yaodong Du, Rania Almajalid, Juan Shan , and Ming Zhang	The four machine learning methods used are artificial neural network (ANN), support vector machine (SVM), random forest, and naive Bayes	Most optimal feature set is determined principal component analysis (PCA) on the feature space.	0.785
3.	MD.REzaul karim, Jiao Jiao, Till DÖhmen, Michael Cochez, Oya Beyan, Dietrich Rebholz-Schuhmann, and Stefan Decker	Gradient-guided class activation maps (Grad-CAM++) and layer-wise relevance propagation (LRP) to highlight regions that distinguish between different classes	The region of interest (ROIs) are extracted using U-Net architecture with ResNet as the backbone.	0.9145

Table 1: Previous Works

Deep learning approaches based on various types of data have been employed in multiple research to identify a variety of issues [6,7,8,9,10]. These include research projects using a variety of techniques and datasets to detect or predict OA using machine learning or deep learning. ANN version for osteoarthritic bone fracture chance classification and a decision-tree model for the prediction of antibody-mediated kidney transplant rejection have been proposed by shaikhina et al.[11]. Despite their great accuracy, they only utilized limited datasets (35 bone specimens and 80 kidney transplants). With 25 accessible samples, Kovanova et al[12] .'s NN study of trabecular bone in OA examined 25 samples. Convolution Neural Network (CNN) was used by Antony et al. [13] to asses the severity of knee OA.

They turned parameters on photos of knee OA and trained a dataset using ImageNet. With the aid of 2-D and 3-D X-ray data, Kobashi et al. [14] examined a post-operative knee function prediction model (52 OA patients). A machine learning algorithm's capacity to categorize magnetic resonance imaging (MRI) of 65 OA patients was assessed by Ashinsky et al. [15]. In order to forecast the incidence of knee OA over the next 30 months utilizing a variety of data, including clinical characteristics, questionnaires, biochemical markers, and pictures, Lazzarini et al. [16] suggested an analytical pipeline based on machine learning. 420 hip X-ray images were used by Xue et al. [17] to test the diagnostic efficacy of CNN in identifying hip OA. Tiulpin et al[3] .'s analysis of X-ray images from 3000 individuals included CNN to identify knee OA.

Using NN, SVM, and random forest among other machine learning methods, Du et al. [18] investigated knee MRI characteristics for OA prediction with PCA. Using 986 pictures, Hirvasniemi used machine learning to predict the prevalence of radiographic hip OA. Using knee X-rays, Brahim et al. [19] used machine getting to know (multivariate linear regression) to perceive early oa inside the knee. Most earlier research centered on deep learning or machine learning with image data from sources like MRI or X-rays. To the best of our knowledge, there isn't much research that combines machine learning or deep learning with statistical data for prediction.

By learning to represent the world as a layered hierarchy of concepts, with each concept defined in relation to simpler concepts and more abstract representations computed in terms of less abstract ones, deep learning is a specific type of machine learning that reaches considerable power and flexibility. Despite the fact that connectionist architectures have been around for more than 70 years, modern architectures and graphics processing units (GPUs) have elevated them to the forefront of artificial intelligence. Several well-known deep learning architectures are introduced, including convolutional neural networks, recurrent neural networks (RNNs), long short-term memory/gated recurrent unit (GRU), self-organizing map (SOM), autoencoders (AE) and restricted Boltzman machine (RBM). Deep belief networks (DBN) and deep stacking networks(DSNs) are also described.

A CNN is a multilayer neural network that takes biological cues from the visual brain of animals. The design is especially beneficial for applications that include image processing. Early layers in a deep network identify features (like edges), while later layers reassemble this information into higher-level input qualities. Differnet types of CNN architectures are:

1. GoogleNet

Understanding CNN Architecture and the GoogleNet Model In 2014, Google researchers introduced Google Net (also known as Inception V1) in the research article "Going

Deeper with Convolutions" with the assistance of several universities. This architecture employs methods like global average pooling and 1–1 convolutions in the middle of the architecture. What GoogleNet offers:
In comparison to earlier state-of-the-art architectures like AlexNet and ZF-Net, Google's architecture is quite distinctive. It employs a variety of techniques, including global average pooling and 1-1 convolution, to build deeper architecture.
This architecture includes vanishing and exploding gradients built-in.

1) Vanishing Gradient
The gradients frequently decrease and finally approach zero as the backpropagation algorithm moves from the output layer to the input layer downward (or backward), essentially maintaining the weights of the lower or beginning layers. Because of this, the gradient decline never reaches the ideal state. The vanishing gradients dilemma describes this situation.

2) Exploding Gradient
On the other hand, in some instances, the gradients continue to grow as the backpropagation method advances. The gradient descent then diverges as a result of the very huge weight updates. The exploding gradients problem describes this situation.

2. VGG16

The Visual Geometry Group is known as VGG. A convolutional neural network with 16 layers is called VGG-16. It is utilized for extremely large-scale image identification, and in the 2014 ILSVRC challenge, this model took first and second place in the aforementioned category. The pretrained network can categorize photos into 1000 different object categories, including several animals, a keyboard, a mouse, and a pencil. As a result, the network now has comprehensive feature representations for a range of photos.

3. MobileNet

Mobilenet is a model that filters images using convolution in a similar manner to that used by CNN in the past, but in a different method. In contrast to standard CNN convolution, it makes use of the concepts of depth convolution and point convolution. This makes CNN more effective at predicting images, allowing them to compete in mobile systems as well. We are employing these convolutional methods as our image recognition model because they greatly cut the comparison and recognition times while yet delivering superior results quickly.

4. EfficientNetV2L

A family of image classification models called EfficientNetV2 outperforms earlier works in terms of parameter efficiency and training speed. Our EfficientNetV2M models are scaled up for quicker training and inference speed and are based on EfficientNetV1 and use neural architecture search (NAS) to simultaneously maximize model size and training speed. Mobile inverted bottle net convolution is employed (MBConv).
Few studies, however, have looked into using deep learning techniques to build mathematical models that can predict osteoarthritis. This study intends to examine the accuracy of several deep learning-based forecast models and to use deep learning technology to diagnose the phases of osteoarthritis. Several deep learning classification models, including Google Net, VGG16, MobileNet, and EfficientNetVL, were taken into consideration in order to achieve this goal. Accuracy, f1-score, precision, recall, and a confusion matrix were used to assess the categorization performance.

III. PROPOSED WORK

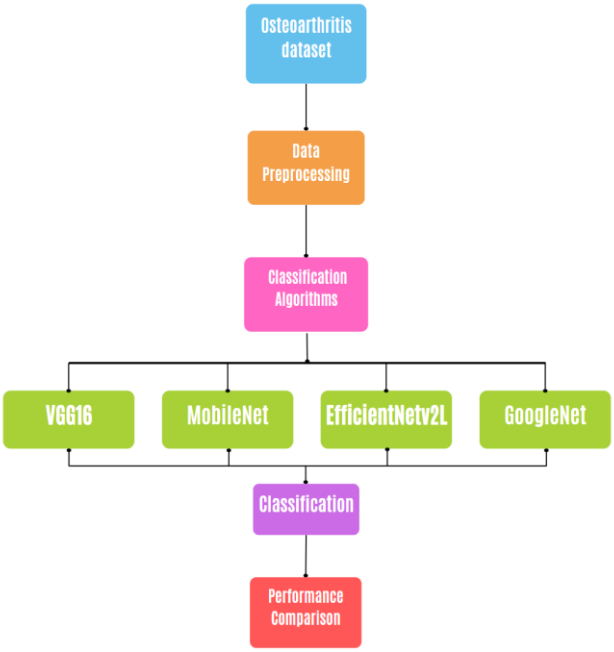


Fig. 2. *System design*

The process flow for detecting osteoarthritis looks like this. The dataset that is crucial for the detection will come first in most cases. Collecting the dataset, followed by the data pre-processing technique to make it noise-free so that better, more accurate results can be obtained, applying various classification algorithms, analyzing each algorithm's performance to see which algorithms are more accurate, and finally performing a comparative analysis of all the algorithms' accuracies are the steps that follow. These results allow us to ultimately draw the conclusion that the patient suffers from osteoarthritis and what kind of treatment is necessary.

Various CNN predefined architectures are used here, including VGG16, MobileNet, EfficientNetV2L, and GoogleNet. Here, we evaluate the dataset's precision and effectiveness across all built-in architectures. Furthermore, depending on our training and testing datasets, these techniques produce varying degrees of accuracy.

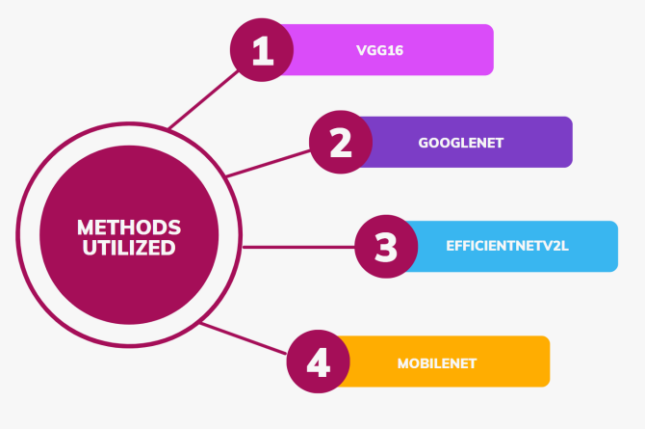


Fig 3. Methodologies

A. TECHNIQUES UTILIZED

1) DATA AUGMENTATION

Data augmentation is the method of creating extra data points from contemporary facts with the purpose of artificially booming the number of records of data. Expanding the data set may require minor tweaks to the data or the use of machine learning models to generate new data points in the latent space of the original data. Many photos that are similar to one another can be produced via data augmentation. Increasing the dataset size in this way helps to decrease overfitting. The cause of this is that as more data is added, the model is unable to overfit all samples and is instead compelled to generalize comparable cutting-edge methods.

Our model essentially applies rotation, horizontal flip, height shift, width shift, and shear to the dataset where noise is added. However, prediction is precise.

2) NORMALIZING

A set of data is transformed to be on a similar scale through normalization. Depending on the data itself, the purpose of machine learning models is typically to recenter and rescale our data so that it is between 0 and 1 or -1 and 1. Normalization is a technique used in data processing to scale the values of numeric columns in a dataset to a common scale when the features in the data have diverse ranges.

The pixel values for RGB images range from 0 to 255. Incorrect values may appear when an algorithm is used. The values of the pixels were translated to the 0–1 range, to obtain clear values.

Several benefits of normalization include:

- Improving training by reducing internal covariate shift
- Reducing network overfitting by promoting regularisation.

3) BATCH NORMALIZATION

The contributions to a layer for each mini-batch are normalized when using the batch normalization technique to train very deep neural networks. As a result, the learning process is stabilized, and the quantity of training epochs needed to train deep neural networks is dramatically reduced.

One possible cause of this issue is that when the weights are refreshed after each mini-batch, the distribution of the inputs to layers below in the network may shift. This can force the learning system to pursue a moving target continuously.

The challenge is that the model is updated layer-by-layer in the opposite order from input to output, with an error estimate that presumes that the weights of the layer preceding the current layer remain constant. Batch normalization offers a comprehensive method for parameterizing practically any deep neural network. Reparameterization greatly reduces the difficulty of revising plans at several levels.

4) L2 REGULARIZATION

Due to the size of the dataset, there is a possibility of overfitting, which renders the model ineffective. Therefore, we employed L2 regularisation to prevent the issue. In L2, you add the data to the model equation to be the sum of the vector's squared times the regularisation parameter's () value, which can be any large number over the amount of the data (m), where (n) is a number of features.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x^{(i)}), y^{(i)}) + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2 \quad (1)$$

5) CATEGORICAL CROSS-ENTROPY

Softmax activation plus a Cross-Entropy loss is categorical cross-entropy, also known as Softmax Loss. We can train a CNN to output a probability over the C classes for each image if we utilize this loss. Multi-class classification is done using it. The gradients become a little bit more complicated when Softmax loss is employed in a multi-label case since the loss has a component for each positive class. Since there are four classes, such as stages 1, 2, 3, and 4, and the categorization is multi-class, categorical cross entropy was utilized.

$$H(p,q) = -\sum_{i} p_i \log q_i = -y\log\hat{y} - (1-y)\log(1-\hat{y}) \quad (2)$$

6) GLOBAL AVERAGE POOLING

A pooling method called global average pooling is intended to take the place of fully connected layers in conventional CNNs. The goal is to produce one feature map in the final mlp (multi-layer perceptron) Conv layer for each category that corresponds to the classification problem. We average each feature map, and the resulting vector is sent straight into the softmax layer, rather than constructing fully connected layers on top of the feature maps. By requiring correspondences between feature maps and categories, global average pooling is more native to the convolution structure than completely connected layers, which is one advantage.

As a result, the feature maps are simply categories of confidence maps. Another benefit of global average pooling is that overfitting is prevented at this layer because there are no parameters to optimize. Additionally, because global

average pooling sums up the spatial data, it is more resistant to input spatial translations.

7) RESIDUAL NETWORK

This architecture introduces the idea of residual blocks to solve the vanishing/explosive gradient problem usi ng a technique known as connection skipping on this network. Skip connections connect layer activations to subsequent layers, bypassing some intermediate layers. This will generate a block of residuals. These remaining blocks are stacked to create resnets.

The strategy behind this network is to let the network fit the residual mapping rather than have layers learn the underlying mapping. Therefore, let the network fit**, F(x):= H(x) - x**, which results in **H(x):= F(x) + x**, be used instead of, say, the original mapping H(x).

B. ACTIVATION FUNCTIONS

1) ReLU

The rectified linear activation function, or ReLU for short, will produce output zero if the input is negative and the input directly if it is positive. Because a model that utilizes it is simpler to train and frequently performs better, it has evolved into the standard activation function for many different kinds of neural networks.

The main advantage of using the ReLU function over other activation functions is that it does not activate all the neurons at the same time

$$RELU(x) = \begin{cases} 0 \; if \; x < 0 \\ x \; if \; x >= 0 \end{cases} \tag{3}$$

2) SOFTMAX

The softmax function is sometimes called the softargmax function, or multi-class logistic regression. This is because the softmax is a generalization of logistic regression that can be used for multi-class classification. Many multi-layer neural networks end in a penultimate layer that outputs real-valued scores that are not conveniently scaled and which may be difficult to work with. Here the softmax is very useful because it converts the scores to a normalized probability distribution, which can be displayed to a user or used as input to other systems. For this reason, it is usual to append a softmax function as the final layer of the neural network. Typically, Softmax has been utilized for multi-class classification due to its effectiveness.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{4}$$

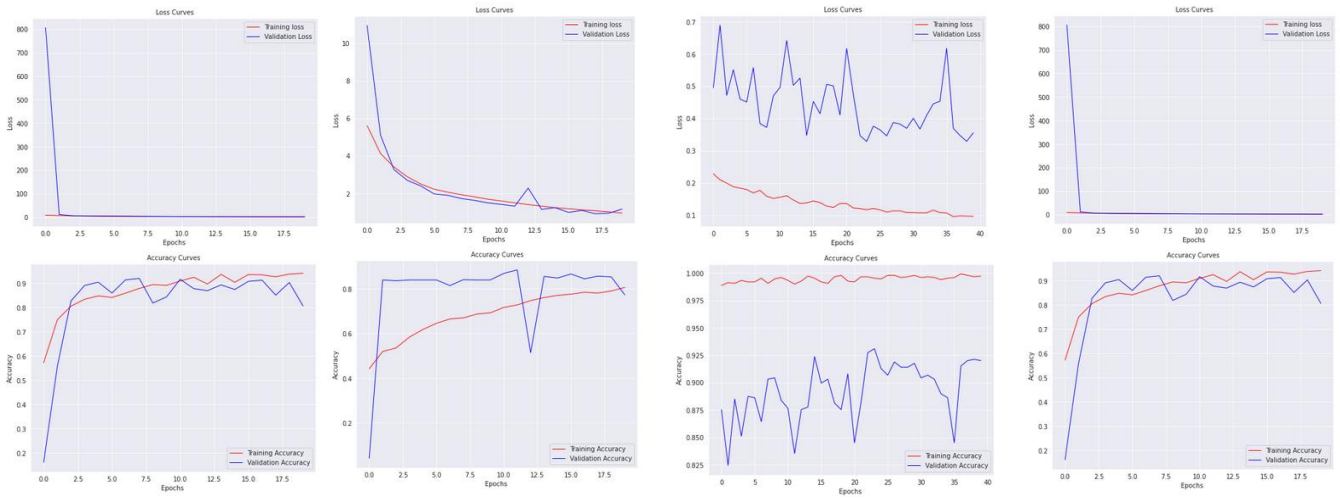
C. HYPERTUNING PARAMETERS

1) MODEL CHECKPOINT	Many applications, like transfer learning and inference utilizing trained models, require the ability to preserve and restore the state of a model. One approach to do this is to store the model's parameters (weights, biases, etc.) in a checkpoint file or directory. This callback offers a few advantages, like: • It chooses whether to save the model at the conclusion of each epoch regardless of performance or keep it just if it has achieved the "best performance" thus far. • It saves the interval. The callback now enables saving after a predetermined number of training batches or at the end of each epoch. • Determines model in its entirety is saved or just the weights are preserved. While running the epochs, the best model is saved using the model checkpoint. The model is the best if one epoch provides the maximum accuracy since it saved that data.
2) DROPOUTS	A neural network's unit discharging is referred to as "dropout." Simply put, a dropout occurs when those randomly chosen neurons are ignored by the units (i.e. neurons). By "not knowing," it meant that these units are not taken into account during a certain forward or backward pass. More specifically, at each level of training, individual nodes 1-p are either removed from the network or kept in with probability p so that the smaller network remains. Dropout enables the deactivation of superfluous neurons, resulting in accurate prediction.
3) ADAM OPTIMIZER	The Adam optimizer combines the following two methods of gradient descent: By taking into account the "exponentially weighted average" of the gradients, this approach is used to speed up the gradient descent algorithm. The technique is used to converge quicker towards the minima when averages are used. Adam is an RMSprop and stochastic gradient descent with a momentum combination. It scales the learning rate using squared gradients, similar to RMSprop and leverages momentum by using the gradient's moving average rather than the gradient itself, similar to SGD with momentum.

4) LEARNING RATE	The learning rate is a hyperparameter that determines how much to alter the model each time the model weights are updated in response to the predicted error. It can be difficult to choose the learning rate since a number that is too little could lead to a lengthy training process that could become stuck, but a value that is too large could lead to learning a suboptimal set of weights too quickly or to an unstable training process. When constructing a neural network, the learning rate is the most crucial hyperparameter. We used a standard learning rate of 0.001 at the start and changed it according to the validation loss that occurred after a few epochs.

Osteoarthritis is a degenerative condition that often causes persistent pain as it progresses over time. Joint stiffness and pain might get bad enough to make going about your everyday business challenging. The pain and limitations of osteoarthritis can lead to depression and sleep problems. CNN architectures are utilized to forecast the patient's OA stage in order to avoid these outcomes.

The following graphs are produced once an examination of osteoarthritis performance is completed.



Google Net *VGG16* *MobileNet* *EfficientNetV2L*

Fig 4. Graphs of Training and Validation Accuracies and Loss of All models

As you can see, there is very little loss in comparison to the epochs, and accuracy improves with each epoch. If you examine the Google Net architecture, here in this study it begins with an input layer, is followed by a functional layer, flattens out, and then adds two different dense layers. Since the architecture of all the models is similar, the differences in accuracy are minimal.

IV. EXPERIMENTAL RESULTS

A. DATASET:

For the purposes of detecting knee joints and grading knee KL, the dataset includes knee X-ray data. The following are the Stage descriptions:

Stage I (Healthy): A healthy knee.
Stage II (Mild): Numerous osteophytes, definite narrowing joint space, and mild sclerosis
Stage III (Severe): Significant joint constriction, large osteophytes, and severe sclerosis.

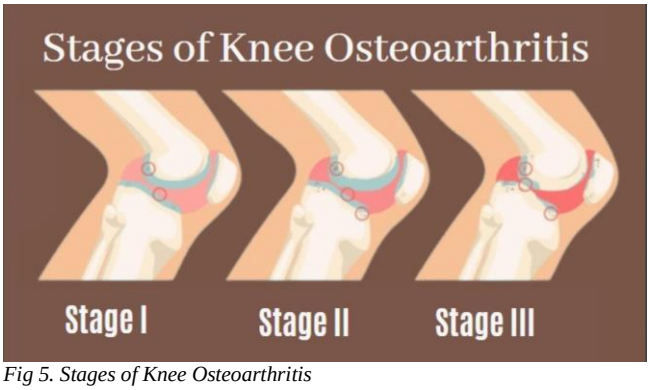


Fig 5. Stages of Knee Osteoarthritis

B. DATA PREPROCESSING

The choice of the appropriate measure is one of the fundamental issues to take into account when dealing with unbalanced data in a classification challenge. Accuracy is frequently employed as the metric, however since accuracy may be deceptive, it would not be a viable choice for the class imbalance issue. The outcomes can be deceptive when a model that was trained on a dataset with the class

imbalance problem is evaluated using accuracy as the criterion. But if we look closely, we can see that the model starts to classify everything as belonging to class 0, which has the effect of giving it a high enough level of accuracy. In these situations, it is advised to integrate additional assessment measures, such as precision, recall, and F1-Score, in addition to using a strategy to address the class imbalance issue. Here we are able to see the outcomes more clearly thanks to the measurements for precision, recall, and F1-Score. Since our data is unbalanced, we balanced it by including images in the training set, validation set, and test set.

A metric of the model's performance across all classes is accuracy. When all classes are equally important, it is helpful. The evaluation process takes into account measures for dataset classification measures like accuracy, sensitivity, specificity, precision, and recall that can be determined by using the confusion matrix. The accuracy is the number of samples properly classified. The proportion of all projected positive samples to the number of positive samples is described by precision. The recall gives information on the proportion of positive samples to all positive samples. These measures are calculated using the confusion matrix's True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values. The following definitions apply to accuracy, sensitivity, specificity, precision, and recall:

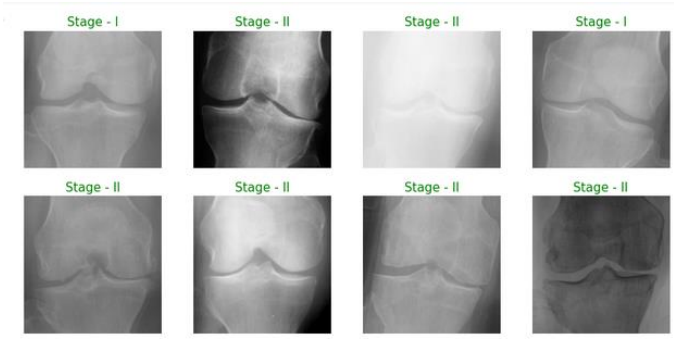
$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (5)$$

$$\text{F1 score} = \frac{(\text{Precision} * \text{Recall}) * 2}{\text{Precision} + \text{Recall}} \quad (6)$$

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (7)$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (8)$$

The classification metric used to assess performance is accuracy.



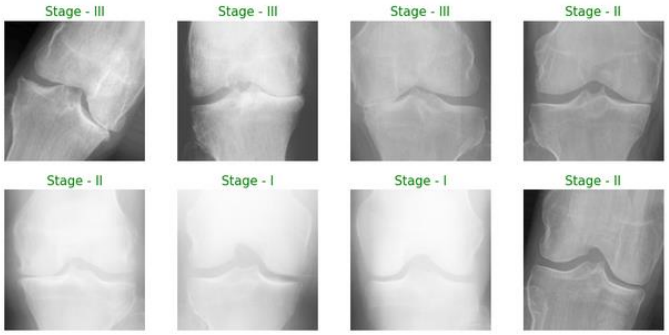


Fig 6. Image of Dataset

C. CLASSIFICATION PERFORMANCE USING DIFFERENT ARCHITECTURE MODELS

We compared four deep-learning algorithms: including VGG16, MobileNet, EfficientNetV2L, and GoogleNet These algorithms were performed in the binary classification case.

"Have you ever had osteoarthritis (OA) diagnosed by a doctor?" this question was utilized to identify patients with OA. Gender, age, region, marital status, education, and household income can be included as independent factors, along with other demographic and personal traits. Following data training, the stages of osteoarthritis (OA) of the knee are predicted.

Architecture	Accuracy	Loss
Google Net	0.9118	1.5342
VGG16	0.8756	0.9529
MobileNet	0.9330	0.0971
EfficientV2L	0.9396	0.9163

Table 2 . Accuracies

The objective of this study is to choose the model that performs the best in terms of classification. Table 2 presents that the EfficientNetV2L model, which has a maximum voting accuracy of 93.96% produced the best results. We also obtained accuracy values of 91.18% for Google Net, 87.56% for VGG16, and 93.30% for MobileNet. These findings demonstrate that, in comparison to other algorithms, the VGG16 model generally produced greater performance.

EfficientNetV2L:

```
Classification Report:
----------------------
              precision    recall  f1-score   support

   Stage - I     0.9749    0.9566    0.9657      1382
  Stage - II     0.7244    0.8251    0.7715       223
 Stage - III     0.8913    0.8039    0.8454        51

    accuracy                         0.9342      1656
   macro avg     0.8635    0.8619    0.8608      1656
weighted avg     0.9386    0.9342    0.9358      1656
```

MobileNet:

```
Classification Report:
----------------------
              precision    recall  f1-score   support

   Stage - I     0.9778    0.9573    0.9675      1382
  Stage - II     0.7344    0.7937    0.7629       223
 Stage - III     0.7258    0.8824    0.7965        51

    accuracy                         0.9330      1656
   macro avg     0.8127    0.8778    0.8423      1656
weighted avg     0.9373    0.9330    0.9347      1656
```

GoogleNet:

```
Classification Report:
----------------------
              precision    recall  f1-score   support

   Stage - I     0.9808    0.9219    0.9504      1382
  Stage - II     0.6316    0.8610    0.7287       223
 Stage - III     0.8302    0.8627    0.8462        51

    accuracy                         0.9118      1656
   macro avg     0.8142    0.8819    0.8417      1656
weighted avg     0.9291    0.9118    0.9173      1656
```

VGG16:

```
Classification Report:
----------------------
              precision    recall  f1-score   support

   Stage - I     0.8831    0.9949    0.9357      1382
  Stage - II     0.8481    0.3004    0.4437       223
 Stage - III     0.4000    0.1569    0.2254        51

    accuracy                         0.8756      1656
   macro avg     0.7104    0.4841    0.5349      1656
weighted avg     0.8635    0.8756    0.8476      1656
```

1) Confusion Matrices

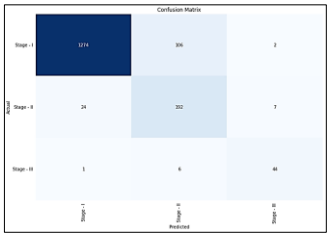


Fig 7. Google Net

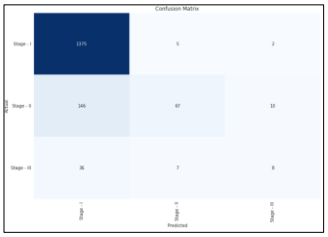


Fig 8. VGG16

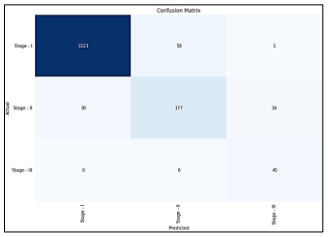


Fig 9. MobileNet

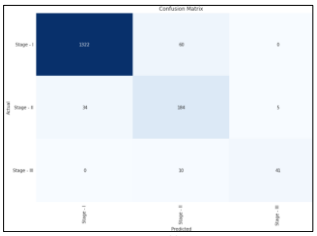


Fig 10. EfficientNet

Comparing all the model's performance

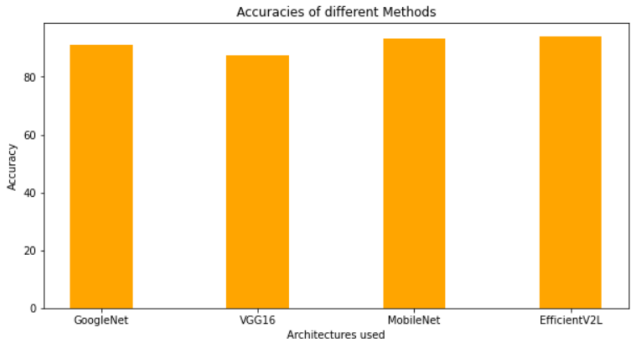
Here, accuracy parameters that are computed for the classification report are compared across all confusion matrices.

V. Conclusion

In this study, we provided a novel representation learning model for X-ray image-based knee OA identification. In general, four prediction models for determining if a person has osteoarthritis were developed and compared. The EfficientNetV2L model has the best classification accuracy, according to the results. This research can help medical professionals swiftly screen for and diagnose osteoarthritis using a wealth of data. The obtained results support previous findings from other studies indicating the medial compartment of the knee contained the most discriminative regions. This study demonstrates that accuracy can still be increased using CNN architectures.

Adding layers promises to improve performance. Each weight is updated so that it moves along the gradient of weight direction, which is the direction along which the loss diminishes. This causes the gradient to gradually decrease, which results in relatively minor modifications to the original layers. As a result, VGG's training period is greatly extended. Given that MobileNet is 32 times more compact than VGG16 while maintaining an accuracy that is at least as high, it must be more effective at acquiring knowledge than VGG. It is true that VGG has far more connections than it requires to carry out its duties. By removing unnecessary connections, VGG16's size can be decreased by a factor of 49 while still providing the same results.

MobileNet is used if model size and inference time are more important than accuracy; otherwise, EfficientNetV2L should be your go-to model, and it can even serve as a straight substitute for MobileNet. EfficientNetV2L is a type of convolutional neural efficiency than previous models, Whereas GoogleNet is a convolutional neural network that is 22 layers deep and has inception modules.



We got an accuracy of 91.18% for Google Net, 87.56% for VGG16, 93.30% for MobileNet, and 93.96% for EfficientV2L. These previously mentioned parameters help us achieve the maximum accuracy for the VGG16 Architecture in this work. Overall, these findings demonstrate that accuracy levels are broadly similar, making it more difficult to select the optimum strategy for the categorization problem.

VI. References

[1] Y. Nasser, R. Jennane, A. Chetouani, E. Lespessailles and M. E. Hassouni, "Discriminative Regularized Auto-Encoder for Early Detection of Knee OsteoArthritis: Data from the Osteoarthritis Initiative," in IEEE Transactions on Medical Imaging, vol. 39, no. 9, pp. 2976-2984, Sept. 2020, doi: 10.1109/TMI.2020.2985861.

[2] The International Workshop on Osteoarthritis Imaging Knee MRI Segmentation Challenge: A Multi-Institute Evaluation and Analysis Framework on a Standardized Dataset Arjun D. Desai, Francesco Caliva, Claudia Iriondo, Aliasghar Mortazi, Sachin Jambawalikar, Ulas Bagci, Mathias Perslev, Christian Igel, Erik B. Dam, Sibaji Gaj, Mingrui Yang, Xiaojuan Li, Cem M. Deniz, Vladimir Juras, Ravinder Regatte, Garry E. Gold, Brian A. Hargreaves, Valentina Pedoia, Akshay S. Chaudhari, on behalf of the IWOAI Segmentation Challenge Writing Group, Naji Khosravan, Drew Torigian, Jutta Ellermann, Mehmet Akcakaya, Radhika Tibrewala, Io Flament, Matthew O'Brien, Sharmila Majumdar, Kunio Nakamura, and Akshay Pai Radiology: Artificial Intelligence 2021 3:3

[3] Tiulpin, A.; Thevenot, J.; Rahtu, E.; Lehenkari, P.; Saarakkala, S. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. Sci. Rep. 2018, 8, 1727.

[4] Progression on MRI Using Machine Learning Methods," in *IEEE Transactions on NanoBioscience*, vol. 17, no. 3, pp. 228-236, July 2018, doi: 10.1109/TNB.2018.2840082.

[5] M. R. Karim et al., "DeepKneeExplainer: Explainable Knee Osteoarthritis Diagnosis From Radiographs and Magnetic Resonance Imaging," in IEEE Access, vol. 9, pp. 39757-39780, 2021, doi: 10.1109/ACCESS.2021.3062493.

[6] Ma, X.; Dai, Z.; He, Z.; Ma, J.; Wang, Y.; Wang, Y. Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction. Sensors 2017, 17, 818.

[7] Izadpanahkakhk, M.; Razavi, S.M.; Taghipour-Gorjikolaie, M.; Zahiri, S.H.; Uncini, A.; Taghipour-Gorjikolaie, M.; Zahiri, S.; Uncini, A. Deep region of interest and feature extraction models for palmprint verification using convolutional neural networks transfer learning. Appl. Sci. 2018, 8, 1210.

[8] Xu, K.; Feng, D.; Mi, H. Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image. Molecules 2017, 22, 2054.

[9] Steven Eyobu, O.; Han, D.S. Feature representation and data augmentation for human activity classification based on wearable imu sensor data using a deep lstm neural network. Sensors 2018, 18, 2892.

[10] Jing, L.; Wang, T.; Zhao, M.; Wang, P. An adaptive multi-sensor data fusion method based on deep convolutional neural networks for fault diagnosis of planetary gearbox. Sensors 2017, 17, 414.

[11] Shaikhina, T.; Lowe, D.; Daga, S.; Briggs, D.; Higgins, R.; Khovanova, N. Machine learning for predictive modelling based on small data in biomedical engineering. IFAC–PapersOnLine 2015, 48, 469–474.

[12] Khovanova, N.A.; Shaikhina, T.; Mallick, K.K. Neural networks for analysis of trabecular bone in osteoarthritis. Bioinspired Biomim. Nanobiomater. 2015, 4, 90–100.

[13] Antony, J.; McGuinness, K.; O'Connor, N.E.; Moran, K. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In Proceedings of the 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 1195–1200.

[14] Kobashi, S.; Hossain, B.; Nii, M.; Kambara, S.; Morooka, T.; Okuno, M.; Yoshiya, S. Prediction of post-operative implanted knee function using machine learning in clinical big data. In Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC), Jeju, Korea, 10–13 July 2016; pp. 195–200.

[15] Ashinsky, B.G.; Bouhrara, M.; Coletta, C.E.; Lehallier, B.; Urish, K.L.; Lin, P.C.; Goldberg, I.G.; Spencer, R.G. Predicting early symptomatic osteoarthritis in the human knee using machine learning classification of magnetic resonance images from the osteoarthritis initiative. J. Orthop. Res. 2017, 35, 2243–2250.

[16] Lazzarini, N.; Runhaar, J.; Bay-Jensen, A.C.; Thudium, C.S.; Bierma-Zeinstra, S.M.A.; Henrotin, Y.; Bacardit, J. A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women. Osteoarthr. Cartil. 2017, 25, 2014–2021.

[17] Xue, Y.; Zhang, R.; Deng, Y.; Chen, K.; Jiang, T. A

preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. PLoS ONE 2017, 12, e0178992.

[18] Du, Y.; Almajalid, R.; Shan, J.; Zhang, M. A novel method to predict knee osteoarthritis progression on MRI using machine learning methods. IEEE Trans. Nanobiosci. 2018, 17, 228–236.

[19] Brahim, A.; Jennane, R.; Riad, R.; Janvier, T.; Khedher, L.; Toumi, H.; Lespessailles, E. A decision support tool for early detection of knee osteoarthritis using x-ray imaging and machine learning: Data from the osteoarthritis initiative. Comput. Med. Imaging Graph. 2019, 73, 11–18.