Global Scientific and Academic Research Journal of Multidisciplinary Studies
ISSN: 2583-4088 (Online)
Frequency: Monthly
Published By GSAR Publishers
Journal Homepage Link- https://gsarpublishers.com/journals-gsarjebm-home/

Fabric Composition Identification using Fine-Tuned Vision Transformers

BY

Chitra G M[1], Shylaja S S[2], Neha Arun Angadi[3], Shreyas Raviprasad[4], Royston Tauro[5]

[1,2,3,4,5]Department of Computer Science & Engineering, PES University, Bengaluru, India"
[1]Research Scholar, VTU, Belagavi, India

Article History

Received: 01/09/2023
Accepted: 07/09/2023
Published: 09/09/2023

Vol – 2 Issue – 9

PP: - 18-25

Abstract

Given the rising trend in retail-based e-commerce from both the producer as well as the consumer, fabric composition identification has gained recent interest. While there are various solutions to this study, they all lack a pragmatic approach and simplicity in design and operation due to the use of NIR sensors and 3D tactile sensors. This research proposes a simple and effective method that requires no extravagant tools and can be tested on photographs taken using a smartphone. To generate a feature vector, a fine-tuned ViT model is employed for feature extraction. PCA-LDA is used to process the features, which are then supplied into an SVM for training and classification. Finally, the generated probabilistic values are calibrated with DGG and used to make predictions. For pure fabrics, the model produced an average F-score of 0.87, log-loss of 0.44, and mean squared error of 0.20. These findings are expanded to include multi-label categorization and composition identification.

Keywords: *Fabric Composition, Vision Transformers, Transfer Learning, PCA-LDA, Probability Calibration, SHAP*

I. INTRODUCTION

The identification of fabric composition is important in various aspects in the industry. Quality assurance in the textile industry is important for both consumers and sellers alike. Consumers would like to be sure of what they have bought and the export rate of any country's textile industry solely depends on its quality.

There has been extensive research in the field of fabric composition identification with the use of image descriptors and filters, near-infrared sensors, gelsight 3D tactile sensors, and optical magnifiers. But these solutions face challenges such as expensive equipment, tedious dataset collection, and controlled environments. Another detail in these solutions is the requirement of data, which is not practical for certain tasks. Even with the advent of data-efficient models, typical image models require at least thousands of images to achieve a decent performance.

This research suggests a transformer model pipe-lined with feature processing and probability calibration for identification of fabric composition to offer a more straightforward solution to these issues. On a pre-trained Vision Transformer model, transfer learning is used due to the magnitude of the dataset. A support vector machine is used to process and train the features, and it produces probabilities by coupling pairs of features. These probabilities are tuned to produce forecasts with greater precision. To generate more calibration samples, utilize data generation. This gets around the requirement for plenty of data.

The interpretability of a model is important to explain the outputs of a machine learning model. Since vision transformers learn self-attention, upon visualizing the attention maps and analyzing SHapley Additive exPlanations plots, we can understand which parts of the image are being given more weightage and attention in the prediction model.

II. INDUSTRIAL RELEVANCE

The industrial relevance of this paper revolves around leveraging the easy availability of smartphones for the purpose of making informative and reliable predictions for the composition of a given fabric material. In an era where convenience is prioritized, the model proposed in this paper uses images of a blended fabric material clicked with smartphones, which can be effortlessly obtained, to predict the composition of the material in question. The application can be observed from two perspectives, that is, seller of clothing in e-commerce and buyer of the same. Different retail business owners can use this model for easy tagging of the material with its composition properties instead of taking a manual approach whereas the buyer will be able to confirm the said composition of the material, hence building the trust for the seller.

*Corresponding Author: **Chitra G M**

III. RELATED WORK

Fabric representation using image processing involves extracting important features from the image. One such method is getting a representation of the textures or textons. Several texture representations exist ranging from gray level co-occurrence matrix, Markov Random Field filters, Gabor Filters, local binary patterns to Convolutional Neural Network based ones like Fusion View - Convolutional Neural Networks. [1] surveyed multiple representation methods and the advancements of them throughout the decade. Most texture descriptors are not able to detect real- world textures at a satisfactory level. Convolutional Neural Network-based descriptors in combination with IFV feature encoding perform better in almost all benchmarks. As each feature descriptor has a drawback of its own, combining different feature descriptors may seem fruitful in accounting for the disadvantages. [3] combined various filters and measured the performance each gave. They found combinations of filters to perform better than any single filter.

The authors [4] talk about quality check, the same goal of our research, but in mango grading. Given how manual grading can be labor-intensive, erroneous, and inconsistent, the authors propose a computer- based grading technique which involves image processing, random forest classifiers, and K-Means clustering to perform final grading using a formula that combines the quality ratings awarded to parameters based on projected categories. They achieved 88.88% accuracy with this grading based on formula after applying image processing techniques. An SVM and CNN comparison can be analyzed as done by [5] where the authors proposed a model for classifying a Mung leaf to check if it is healthy or has a disease. Different architectures for SVM and CNN were prepared and trained to compare the performance on the complex features extracted for various diseases. The study showed CNN to have a better performance at 95.05% of identification accuracy.

Evolving from the texture descriptors to Convolutional Neural Network-based ones, many models were brought forward to perform fabric classification. [2] used ensemble convnets to tackle the problem of multi-class classification. They make use of the Xception model for feature extraction and several tail Convolutional Neural Networks with a sigmoid activation for each class. Convolutional Neural Network-based models tend to be more computationally expensive and complex. One way to go about it is looking at the filters in each convolution layer as a set of filter banks. [6] introduced the Texture-Convolutional Neural Network (T-CNN) which used the concept of energy and pooling activation outputs of each layer. They were able to reduce the complexity and computation time. [7] used a wavelet scattering network (ScatNet) which helped give stable translation invariant image representations. The first layer outputs Scale-Invariant Feature Transform-type descriptors which can be improved upon by using wavelet scattering vectors.

There were several approaches to combine the strengths of feature descriptors and deep networks. [8] introduced NmzNet, a handcrafted convolution network which used ScatNet and a Gabor filter bank to compute convolutions in each layer. Fisher vector aggregations were used to get the final discriminative features. NmzNet performed better in most benchmarks if not at par with state-of-the-art models. [9] integrated Local Binary Patterns and Gabor layers into a capsule network to get better feature extraction. [10] combined global features from ScatNet and Local Binary Patterns for texture representation. While the results are at par on experimental databases, their features are more distinct and robust. [21] introduced Ensemble networks to classify fabrics i.e., have a head layer and one single convolution neural network for each class and then get the max value for the correct classification.

[20] introduced zero-shot learning to classify fabrics, they focused on more material aspects, and used gelsight sensor-based images to classify fabrics. Using Gelsight sensors-based images performed better than the rest with them getting an accuracy of 90%. [19] proposed Optical Coherence Tomography images-based classification, they found the composition for 3 fabrics i.e., cotton, wool, and polyester and they were quite successful with it.

An analysis of a pre-trained transformer model helps understand how a pre-trained model has contributed in the prediction-based models. The purpose of [17] is to solve image processing issues using a pre-trained transformer model (IPT). The pre-trained transformer model (IPT) is built with several heads, multiple tails, and a common transformer body to serve various image processing needs. As researched by [18], transformer models have shown to be effective pre-training frameworks for a variety of natural language processing applications. The effectiveness of transformers in natural language processing lays the groundwork for future research into the benefits and power of transformers in computer vision and image processing. They do away totally with recurrence and convolutions in favor of a fresh, straightforward network design that relies only on an attention mechanism. Studies on two machine translation tasks reveal that these models are more parallelizable, produce better quality results, and take a lot less time to train.

Multi-label classification is used to predict the composition of blended fabrics. There are several algorithms that handle multiple labels and can be used for fabric composition as well. [12] introduced CU-Net which independently models the composition for each class using a restraining loss. They use deep feature extraction and component unmixing to get accurate predictions on the composition. Fabric composition can be obtained using near-infrared sensors which use the spectrum of reflected light from the fabric. The peaks in the wavelengths help determine the different blends and their composition. [13] introduced FabriTell which uses near-infrared sensors and analysis techniques to predict the composition of fabrics. It supports 17 two-component blends of common materials like cotton, polyester, etc.

Once the features are extracted, they are used to classify the

*Corresponding Author: **Chitra G M**

images of fabrics into different classes. Some scenarios exist where a single sample belongs to multiple classes with some proportional amount. Such is the case with fabric composition. The probability scores given by a support vector machine do not accurately represent the composition of the fabrics and must be calibrated. [11] generated data to help calibrate for small datasets. The calibration is done using isotonic regression and Naive Bayes is used to generate the new data. This research observed that isotonic regression calibration is difficult with small data sets, and using it could make the calibration of unobserved data worse. By producing additional calibration data, it is possible to solve the issue by making greater use of the information in the data.

IV. PROPOSED METHOD

The availability of a limited dataset and the calibration of probabilities for multi-label classification are the two key issues that need to be handled in this situation. Since it was challenging to get a sizable, high-quality fabric dataset, no deep learning network could be trained due to the problem of overfitting and would hence be useless. Transfer learning is the greatest choice when taking into account these considerations. Considering this, a pipeline, as elaborated in Fig. 1, is suggested that comprises of a visual transformer, feature post-processing, support vector machine classification, and probability calibration.

A. *Transformer Model*

Self-attention is used by vision transformers to discriminate between significant and non-significant elements in the input. This is accomplished by dividing the embeddings into the three components of the query, key, and value vectors. To determine the score for each embedding in the input, the query and key vectors are employed.

a. $score = Q.K^T$

The scores are then normalized by dividing them by the square root of the dimension of the vectors. This is done to stabilize the gradients during training. The value vector is then multiplied with the softmax of this score to obtain the self-attention.

b. $softmax\left(\frac{QK^T}{\sqrt{d}}\right)V$

This self-awareness is subsequently transferred to a feed-forward neural network. It is common knowledge that vision transformers can capture connections between various regions in the image. We are able to capture many facets of the image thanks to our self-attention. A Vision Transformer model pre-trained on ImageNet-21k ([15]) at resolution 224x224 was chosen for this experiment.

The 224x224 Vision Transformer model ([14]) from Hugging Face was used for the purpose of this research. Further inspection showed that the model uses 12 layers with 12 attention heads each. The model breaks the resized image into 49 patches of 32x32. Since the dataset contained 4 classes, a final fully connected layer of size 4 was used to fine-tune the model. All the layers were frozen and only the last layer was allowed to train on the data. The

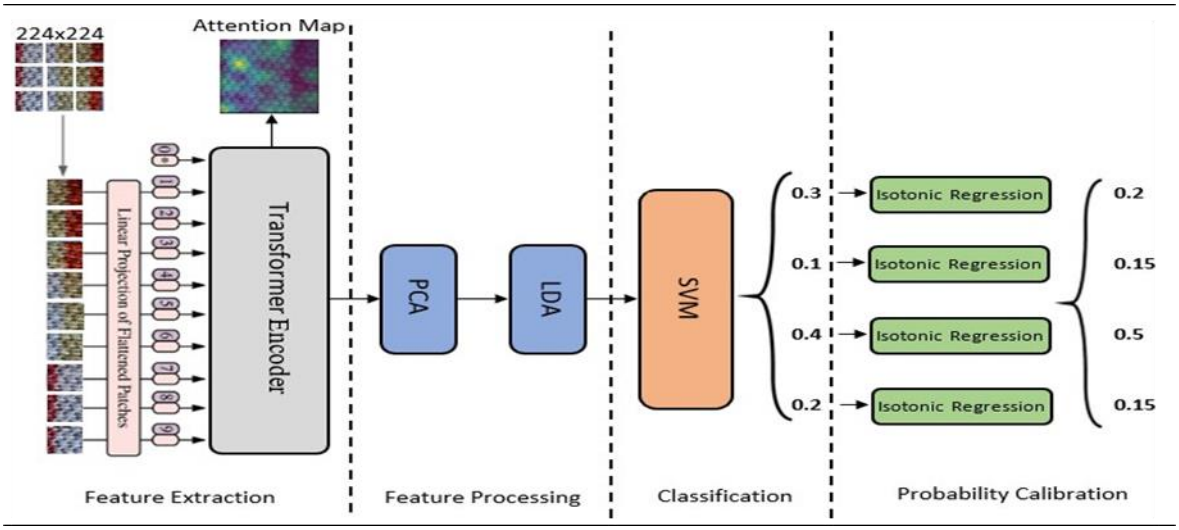

Fig. 1: Pipeline of the proposed model feature vector was taken as the output of the last layer giving a dimension of 200.

A. *PCA-LDA*

Data in a high- dimension space can be tricky to work with. Due to high dimensionality, data points become really sparse and can be bad for classification, especially when trying to predict class probabilities. This is known as the curse of dimensionality.

Principal Component Analysis was used to reduce the dimensionality of the features. The number of components or an explained variance can be specified for Principal Component Analysis. Explained variance measures the proportion to which the model accounts for the variation in the data. A higher explained variance means less loss of data and a lower value means smaller dimensions. A typical value of this tradeoff is 0.99 which is what was used here as well.

Linear Discriminant Analysis is a supervised classification technique. It also serves as a dimensionality reduction technique. It tries to maximize the inter-class variance and minimize intra-class variance. Let μ_1 and μ_2 be the means of two clusters and s_1 and s_2 their respective variances. Linear Discriminant Analysis optimizes the following function in

*Corresponding Author: **Chitra G M**

order to achieve this:

c. $\frac{(\mu_1-\mu_2)^2}{s_1^2+s_2^2}$

Linear Discriminant Analysis performed on 4 classes gives a final feature vector of dimension 3.

B. Support Vector Machine

Support Vector Machine is a powerful classification algorithm that boils down into an optimization subject to some class constraints. Another reason it works well for high-dimensional data is due to its kernel trick. The Radial basis function kernel was used with a penalty parameter C of 1. The Support Vector Machine uses Platt scaling to calculate probabilities in the case of binary classification. It uses pairwise coupling for multiclass classification. The probabilities given by the Support Vector Machine can't be completely depended on as it isn't calibrated.

C. Probability Calibration

Probability calibration is a necessary step as probabilities from the SVM aren't accurate. Plotting a calibration plot which maps the predicted value to the fraction of true positives gives an idea of how well the model is predicting. A straight line means a perfect model. To measure this quantitatively, The Brier score loss is used which measures the mean squared difference between the predicted values and the actual outcomes. The smaller the value the more calibrated the model is. Given the dataset size, typical calibration cannot be performed efficiently. Calibration requires a lot of data and becomes inefficient for small datasets. To calibrate small datasets, Data

Fig. 2: Samples from the dataset

Generation was used to synthetically create 3000 samples which is enough to perform the calibration. Data Generation is done using a Support Vector Machine. A 70-30 split suggested by [11] was used. Data was generated for all classes in the same manner and an isotonic regression model was trained on these classes individually giving 4 models in total. Each model is capable of mapping the predicted probability scores from the SVM to a calibrated one for each class.

I. DATASET

During the course of all the experiments, the dataset went through three iterations in order to obtain to the final dataset used for training the model. The initial dataset consisted of 627 images taken on a smartphone. All images were taken under approximately the same height and lighting conditions. Classes included cotton (266), silk (135), synthetic (118), and polyester (108). All images were of size 3472x4624 of both portrait and landscape orientation.

Despite being large, the aforementioned dataset lacked in detailing and had a significant bias towards designs present on the cloth. To address this issue, dataset was collected among various sources which consists of images taken using a 50mm macro lens with a zoom of 3-5cm. Some images were taken from the Fabrics dataset ([16]), mainly for classes which were more difficult to obtain. The final dataset contains 937 images and classes included cotton (245), Nylon (132), Silk (261), and Polyester (167).

This dataset comparatively gave better results as it consisted of images that were more detailed and removed all bias of color, patterns, designs on the cloth, and other info which would add extra bias. Fig. 2 shows some samples of the dataset.

V. RESULTS

A. Transformer Model

A pre-trained model ([22]) was used for fine-tuning. It was trained for 20 epochs on the dataset

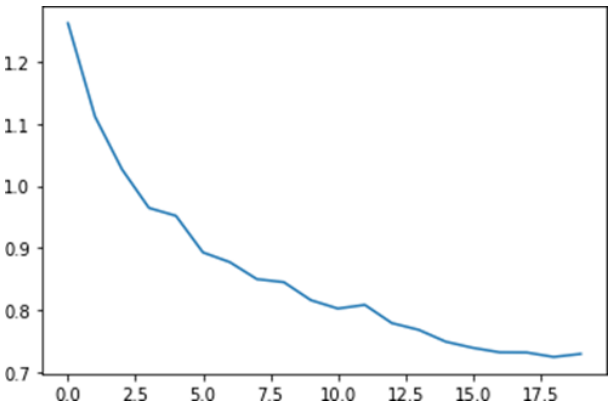

Fig. 3: Cross entropy loss vs no. of epochs with the help of Adam optimizer with a learning rate of 0.001 and Cross Entropy loss:

d. $$-\sum_{x=1}^{c} p(x)\log q(x)$$

where $p(x)$ is the true class distribution and $q(x)$ is the prediction of the model. It achieved an accuracy of 26.38% and a loss of 0.72. Since the model is being used as a feature extractor, the accuracy was not a concern rather the distinction of features between classes. t-SNE was used to visualize all the features.

As can be seen from Fig. 4, fine-tuning the model gave better feature representation in terms of clear distinction between classes.

A. PCA-LDA

Principal Component Analysis was used as a dimensionality reduction technique. Instead of using a fixed value of the number of components to use, explained variance was used. An explained variance of 0.99 was used and the variance was plotted against the number of components. 81 dimensions was

found to be the number of components to account for most of the variation in the data.

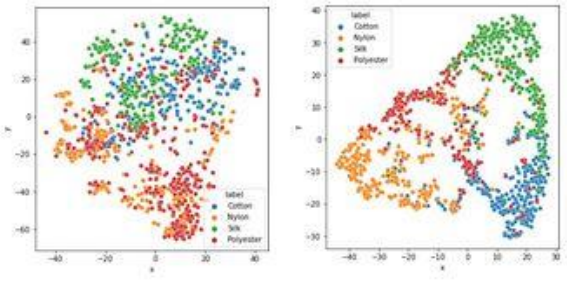

Fig. 4: Features extracted from the pre-trained ViT (left) and features extracted from the fine-tuned model (right)

Linear Discriminant Analysis was further applied on the feature vector obtained from Principal Component Analysis. Sklearn had eigendecomposition and automatic shrinkage using the Ledoit-Wolf lemma. It was trained on 80% of the data and gave a final dimension of 3. As can be seen from Fig. 5, Linear Discriminant Analysis proved to provide more distinction between classes for classification.

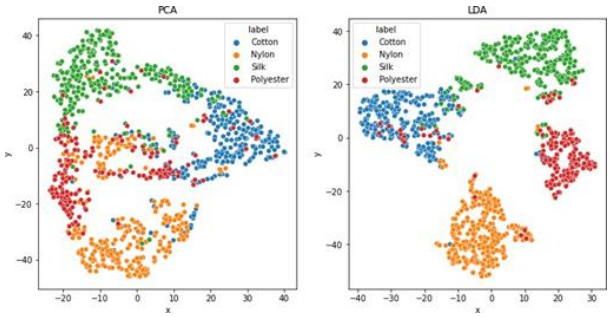

Fig. 5: Features extracted from Principal Component Analysis (left) and Linear Discriminant Analysis (right)

A. Support Vector Machine

A Support Vector Machine was trained on the resulting features obtained from Linear Discriminant Analysis. A radial basis function kernel with a penalty parameter C of 1 was used. Probabilities from the Support Vector Machine were used to predict the actual probabilities for each class. The model gave an average accuracy of 87%, a log loss of 0.44, and mean squared error of 0.20.

TABLE I: SVM Performance

Fabric Label	Classification Metrics		
	Precision	***Recall***	***Fscore***
Cotton	0.846154	0.830189	0.838095
Nylon	0.918033	0.949153	0.933333
Silk	0.820000	0.931818	0.872340
Polyester	0.920000	0.718750	0.807018

A. Probability Calibration

The calibration of probabilities was observed to not give any significant boost in performance. The log loss and mean squared error remained about the same while the average f-score dropped to 0.83. The f-score of cotton improved by 1% while all the other f1-scores were noted to have dropped by some amount.

The composition was tested with 33 test images and the metrics used were mean squared error and mean average error. The calibrated values achieved a mean squared error of 0.157 and a mean average error of 0.276.

TABLE II: Composition predictions

Ground Truth	Predicted composition			
	Cotton	***Polyester***	***Nylon***	***Silk***
Polyester + Acrylic blend	0%	2%	95%	1%
Polyester + Nylon blend	0%	45%	50%	5%
60% Cotton 40% Polyester	45%	45%	8%	0%
52% Cotton 48% Polyester	42%	4%	20%	31 %
Cotton + Nylon blend	57%	36%	4%	1%
80% Cotton 20% Polyester	53%	5%	1%	39 %
72% Cotton 28% Polyester	46%	29%	22%	1%
35% Cotton 65% Polyester	2%	6%	0%	91 %

DISCUSSION

Initial experiments for the model were conducted using image descriptors and filters such as Histogram of Oriented Gradients, Local Binary Pattern, gray level co-occurrence matrix, Scale Invariant and Feature Transform, Speeded Up Robust Features, etc. The experiments showed a consistently poor performance from all these descriptors. Convolutional Neural Networks and vision transformers alone proved to perform better and when combined with different filters, performed about the same with no significant improvement in performance. It also wasn't clear whether the performance was due to the deep neural networks or the image descriptors, and how much each contributed.

Transfer learning on deep neural networks proves to be an effective way of extracting features without having to worry about overfitting problems and training time. This can be seen from the way the fine-tuned model extracted distinct features. Visualizing the attention maps would help understand where the model was looking for important features. It was observed that the attention drifted off towards the edges of the image at the deeper layers as can be seen in Fig.8. This may be due to the types images the Vision Transformer was trained on. Fabrics images do not have features that resemble that of other classes such as cats, dogs, or cars. This may mean that more fine-tuning is required or domain adaptation methods

*Corresponding Author: **Chitra G M**

need to be used.

While probability calibration provided a small improvement in log loss, other methods can be explored that better suit the problem at hand. The probabilities can be calculated by other means that take into account the characteristics of individual

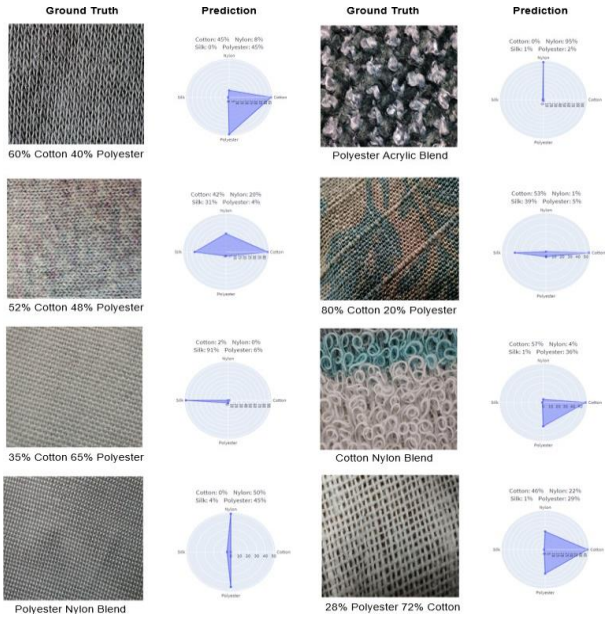

Fig. 6: Model predictions fabrics.
This may give probabilities that are independent of each other.

SHapley Additive exPlanations

Model explainability is an important factor in trying to understand the predictions of machine learning models. SHapley Additive exPlanations (SHAP) was used to bring more under-standing to the model's predictions. Shapley value corresponds to how much influence an input feature has on the overall output of the model. It is calculated using coalitions of features and sampling them from the data and changing one feature to get different combinations of these features.

For data in the format of an image, the input features are the pixels and it can become really large and expensive to compute all combinations, due to which they are approximated. SHapley Additive exPlanations shows us which parts of the image contribute to each class both positively and negatively.

A higher value suggests a stronger influence of that part of the image on the output. This way, we can infer which features of the image might help the model attribute the image to a certain class.

A max evaluation of 1000 was used with a batch size of 50 to obtain the SHapley Additive exPlanations plots. A higher number of evaluations means more fine-grained plots but it is comparatively more expensive. Thus, a balance must be ensured between these two factors. The plots show the classes in decreasing order of probabilities.

While the SHapley Additive exPlanations plots showed parts of the image that contributed towards a particular class, one thing looked to be consistent. The SHapley Additive exPlanations plots of silk images are finer-grained than the other classes. This seems to be consistent with silk cloths whose threads are more finely woven than those of cotton and polyester. This also shows in the high recall of silk images

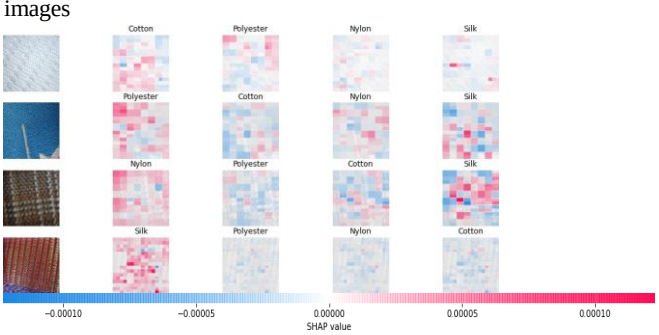

Fig. 7: SHapley Additive exPlanations plots
For some sample data and also how distinct the samples are in the feature space of Linear Discriminant Analysis.

This was a valuable insight into this field as finding the composition of fabrics via images was previously done without giving a clear description of why it happened. SHapley Additive exPlanations plots help us describe the hidden blackbox which Convolutional Neural Networks introduce into itself, hence making it more descriptive and understandable.

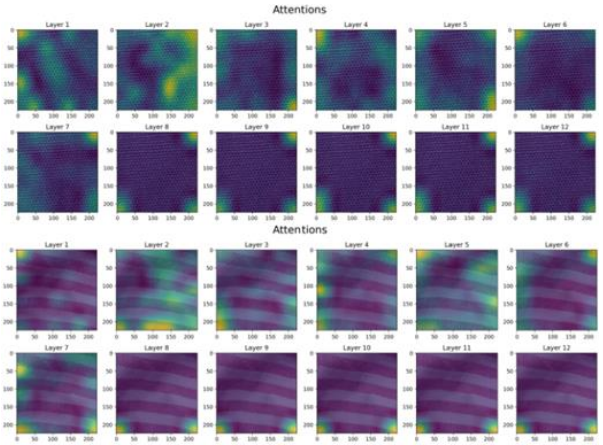

Fig. 8: Attention maps of sample images CONCLUSION

In this paper, we developed an approach that predicts the composition of the fabric without the need for a large dataset. We make use of vision transformers which implicitly break the image into patches, which help detect repeating patterns and fiber characteristics. Fine-tuning the Vision Transformer on our dataset helps extract features giving better distinction between classes. The probabilities from the Support Vector Machine are calibrated to achieve better a representative of the composition and SHapley Additive exPlanations plots are used to get a better understanding of the model's predictions along with visualizing the attention maps of the transformer. Better feature extraction methods can be debated to be implemented in order to improve the proposed model since the features are the one of the core factors in the training to make predictions. Multi-label classification is another field that can be explored since adding

*Corresponding Author: **Chitra G M**

in more fabrics as a part of core dataset for the prediction model can expand the scope for more blended fabrics that are commonly available. Probability calibration is another domain to be further researched in since different calibration techniques have different focus points for determining the final calibrated probabilistic values.

ACKNOWLEDGMENT

We want to express our thanks to Dr. Hanumantha Naik H S, Assistant Professor in the Department of Textile at Sri Krishna Rajendra Silver Jubilee Technological Institute. His insights and contacts helped us collect the data we required for our research.

We are grateful to Dr. Naveen Padaki, Scientist at the Central Silk Technological Research Institute, Central Silk Board. He was patient in helping us and giving critical feedback on our project. To further enhance our dataset, we were able to gather data at the Central Silk Board.

The dataset would not have been complete without the help of Mr. Radhakrishna from the Textile Committee, Bangalore. He was diligent in providing us with the composition of all the samples and their ground truth.

We would like to thank Ms. Nithya M P who helped take the photos of the initial dataset and Uchit, who helped label the data and organise everything into corresponding folders for easier access.

REFERENCES

1. Liu, L., Chen, J., Fieguth, P. et al. From BoW to CNN: Two Decades of Texture Representation for Texture Classification. Int J Comput Vis 127, 74–109 (2019).
2. Ohi, Abu Quwsar et al. “FabricNet: A Fiber Recognition Architecture Using Ensemble ConvNets.” IEEE Access 9 (2021): 13224-13236.
3. Barley, A. and Town, C. (2014) Combinations of Feature Descriptors for Texture Image Classification. Journal of Data Analysis and Information Processing, 2, 67-76.
4. D. Supekar & M. Wakode, “Multi-Parameter Based Mango Grading Using Image Processing and Machine Learning Techniques”, INFOCOMP Journal of Computer Science, vol. 19, no. 2, pp. 175–187, Dec. 2020.
5. Naik & D. H. T. Thaker, “Early Recognition of Mung Leaf Diseases based on Support Vector Machine and Convolutional Neural Network”, INFOCOMP Journal of Computer Science, vol. 21, no. 1, Jun. 2022.
6. Andrearczyk, Vincent & Whelan, Paul. (2016). Using Filter Banks in Convolutional Neural Networks for Texture Classification. Pattern Recognition Letters. 84. 63-69.
7. J. Bruna and S. Mallat,” Invariant Scattering Convolution Networks” in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 35, no. 08, pp. 1872-1886, 2013.
8. N. -S. Vu, V. -L. Nguyen and P. -H. Gosselin,” A Handcrafted Normalized-Convolution Network for Texture Classification,” 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 1238-1245.
9. Patrick, Mensah & Amponsah, Anokye & Agyemang, Kwame Baffour & Armah, Gabriel & Ayidzoe, Mighty & Umar Bawah, Faiza & Adekoya, Adebayor & Weyori, Benjamin & Amo-Boateng, Mark. (2021). Multi-lane LBP-Gabor Capsule Network with K-means Routing for Medical Image Analysis. International Journal of Advanced Computer Science and Applications. 12.
10. Vu-Lam Nguyen, Ngoc-Son Vu, Hai-Hong Phan, Philippe- Henri Gosselin. LBP-and-ScatNet-based Combined Features For Efficient Texture Classification. Multimedia Tools and Applications, Springer Verlag, In press, ff10.1007/s11042-017-4824-5ff.
11. Alasalmi, Tuomo & Koskima¨ki, Heli & Suutala, Jaakko & Ro¨ning, Juha. (2018). Getting More Out of Small Data Sets - Improving the Calibration Performance of Isotonic Regression by Generating More Data.
12. Feng, Zunlei & Liang, Weixin & Tao, Daocheng & Sun, Li & Zeng, Anxiang & Song, Mingli. (2019). CU-Net: Com- ponent Unmixing Network for Textile Fiber Identification. International Journal of Computer Vision. 127. https://matoha.com/fabrics-identification
13. Wu, B., Xu, C., Dai, X.,Wan, A.,Zhang, P.,Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., & Vajda, P. (2020). Visual Transformers: Token-based Image Representation and Processing for Computer Vision.
14. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248–255).
15. Kampouris, S. Zafeiriou, A. Ghosh, S. Malassiotis, Fine- grained material classification using micro-geometry and reflectance, 14th European Conference on Computer Vision, Amsterdam, 2016
16. Chen, Hanting & Wang, Yunhe & Guo, Tianyu & Xu, Chang & Deng, Yiping & Liu, Zhenhua & Ma, Siwei & Xu, Chunjing & Xu, Chao & Gao, Wen. (2021). Pre-Trained Image Processing Transformer. 12294-12305. 10.1109/CVPR46437.2021.01212.
17. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia
18. Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.
19. Sabuncu, Metin&O¨ zdemir, Hakan.(2021).Classification of Material Type from Optical Coherence Tomography Images Using Deep

*Corresponding Author: **Chitra G M**

Learning. International Journal of Optics. 2021. 10.1155/2021/2520679.

20. F.Wang, H. Liu, F. Sun and H. Pan, “Fabric recognition using zero-shot learning,” in Tsinghua Science and Technology, vol. 24, no. 6, pp. 645-653, Dec. 2019, doi: 10.26599/TST.2018.9010095.
21. Ohi, Abu & Ph. D., M. & Hamid, Md. Abdul & Monowar, Muhammad Mostafa & Kateb, Faris. (2021). FabricNet: A Fiber Recognition Architecture Using Ensemble ConvNets.
22. https://huggingface.co/google/vit-base-patch16-224-in21k