Predicting the Business Domain of Companies with Logos

Abstract— **In this project, our aim is to know if there are implicit visual logo features that are common across businesses within a domain. To demonstrate this, we are attempting to predict the business domain given a business logo. We start from raw image embeddings, then use more sophisticated methods to improve feature localization, thereby improving the quality of classification. Finally, an interesting result we obtained showed that the text inside the image predicts business domains with much better accuracy than just the logo itself.**

I. INTRODUCTION

Business logos are considered the face of the business and industries, serving as a representation of the identity and values of their brand. These logos play an important role in making a good first impression on customers and stakeholders. A well-designed logo can convey reliability and uniqueness, all of which are important for building a brand's presence in each sector.

The importance of business logos extends beyond the visual elements; they act as an important feature for marketing strategies, brand recognition, and firm performance. Logos create a consistent image across various platforms, from digital to physical products, and build familiarity with customers. Over time, logos have become a symbol of trust and loyalty, reproducing deeply with the audience's perception of the brand. [16].

Despite their critical role in branding, the specific design elements that distinguish logos across different industries are not well understood. Each industry often follows unique design conventions, such as specific color palettes, typography, or other features, which align with its target audience. analyzing these elements can provide valuable insights into the implicit patterns that define industry-specific branding. This understanding has the potential to refine logo design practices and enhance their effectiveness in communicating a brand's identity.

Furthermore, advancements in machine learning have opened new avenues for exploring the relationship between logo design and business domains. By using image analysis and predictive modeling techniques, researchers can identify the visual and textual features that distinguish logos across industries. This approach not only helps to understand existing trends but also offers a framework for creating more impactful logo designs tailored to the demands of specific sectors.

II. RELATED WORK

Convolutional Neural Networks and Vision Transformer Models have been used by other researchers to identify and classify logos. This survey [20] conducts a comparative analysis of different types of deep learning models for logo classification and reports that CNNs have an advantage over Vision Transformers in terms of real-time responsiveness and privacy constraints.

The visual features of a logo play an important role in establishing the brand identity and shaping the initial perception of a company's trustworthiness. [17] mentions that visual elements in logos are processed faster by the brain compared to text, making them more memorable. It also highlights that the color, font, and overall composition of a logo are key factors in conveying a company's value proposition. Poor font or color choices could lead to misinterpretation or distrust from potential customers.

During initial analysis, many logos appeared to contain significant background whitespace. While it may seem intuitive to remove the whitespace and focus only on the foreground elements as relevant data, [1] suggests that machine learning models often rely on spurious features like background whitespace for predictions. Using the CelebA dataset, they demonstrate that removing spurious features can benefit some classification

groups but negatively impact others, even in balanced datasets. They conclude that removing spurious features can often reduce accuracy, even when core features accurately define the targets.

Mixup [3] is a modern image augmentation technique designed to improve model robustness and generalization by creating diverse training samples. It generates virtual examples by linearly interpolating two random images from the training set and their labels, using a mixing ratio (λ) sampled from a beta distribution. Mixup has been shown to improve generalization error in state-of-the-art image models on datasets like CIFAR and ImageNet. Similarly, **CutMix** [4] is another advanced augmentation technique that cuts rectangular patches from one image and pastes them onto another image. The labels of the two images are mixed proportionally based on the area of the patches. Like Mixup, CutMix uses a mixing ratio (λ) sampled from a beta distribution. Unlike Mixup, CutMix avoids information loss by retaining informative pixels, forcing the model to learn from partial views of images. CutMix has consistently outperformed state-of-the-art augmentation strategies on datasets like CIFAR and ImageNet.

Gradient-weighted Class Activation Mapping (Grad-CAM) [12] is a visualization technique used to produce coarse localization maps that highlight important regions in an image for predicting specific labels. Grad-CAM uses gradients flowing into the final convolutional layer to create feature maps. We used Grad-CAM to visualize how models process logos and identify which areas certain classes focus on within a logo.

Many company logos also contain text, often including the company name, slogans, or regional information. This text can be valuable for identifying the industry of a company. **EasyOCR** [13], a Python library compatible with PyTorch, achieves high text recognition accuracy (95%) using CRAFT models for text detection and Convolutional Recurrent Neural Networks [5] (CRNN) for text recognition. Similarly, **Tesseract OCR** [19] is another widely used OCR tool for tasks such as digitizing books or processing forms. [18] compares EasyOCR and Tesseract OCR, reporting that EasyOCR achieved 95% accuracy compared to Tesseract's 90%.

Company logos consist of multiple colors, and these colors often represent the domain of the company. Therefore, extracting colors from company logos and analyzing them can help identify certain trends in color usage. [14] compares the performance of RGB and HSV color segmentation models on road signs. The paper concludes that the HSV model performs better at color detection compared to the RGB model, as the HSV model is more robust to lighting changes. Similarly, [15] derives comparable results and emphasizes that the Value component in the HSV model plays an important role in distinguishing intensity using Hue or Saturation information.

III. Datasets

A. Datasets Explored

CrunchBase Business Logos Dataset: Initially, we explored the CrunchBase Business Logos dataset; however, most logo image URLs were broken, making direct retrieval difficult. Furthermore, the dataset primarily included startups, whose logos and public financial data were not easily accessible online. For our analysis, we aimed to focus on more prominent companies with distinctive logo features and accessible metadata. Consequently, we decided not to use this dataset.

7+ Million Company Dataset on Kaggle by People Data Labs: Next, we explored a comprehensive dataset from People Data Labs [11] on Kaggle, which contained data for over 7 million companies. This dataset included crucial information such as company names, industry categorization, country and city of origin, and estimated employee count. The estimated employee count provided a useful gauge of a company's popularity, enabling us to identify and prioritize logos from larger, more well-known companies. This focus on companies with sizable workforces served as a proxy for popularity, aligning with our objective to analyze the visual logo features of more recognizable businesses.

Scraped Custom Image Dataset for Logos: Using the company names obtained from the People Data Labs dataset, we wrote a script to fetch corresponding logos via the DuckDuckGo image search API. By automating this process, we gathered logo images for companies with complete

information, including name, industry, and location. This method enabled us to build a collection of logos linked to key attributes such as country of origin and industry category.

B. Method of Acquisition

First, we aimed to compile a list of popular company names. We obtained over 7 million company names from the Kaggle dataset, which also included an estimated number of employees for each company. This employee count was used as a proxy for popularity.

To reduce the scope of the analysis, the dataset with 7+ million records was sampled to ensure a diverse pool of industries with varying properties while keeping it manageable for exploration, acquisition, and building predictive models. The most popular industries were ranked in descending order based on the average number of employees within each industry. From this ranking, the top 8 industries were selected, and approximately 12,000 of the most popular companies (gauged by employee count) were chosen for each industry. This sampling reduced the initial dataset to 96,000 records for further analysis.

The next step was to associate each company with its logo. Since the dataset only contained company names and other non-visual features, logos had to be scraped from the internet. The company names were used as search queries, and the DuckDuckGo Image Search API was employed to retrieve image URLs corresponding to these names. The logos were then downloaded using the obtained URLs.

Due to rate limits imposed by the API and other errors encountered during the search and download process, we successfully acquired roughly 10,000 logo images for each industry.

Finally, we normalized the company names and their corresponding image paths to ensure compatibility with dataset versioning platforms like Kaggle. This resulted in a dataset of approximately 80,000 company names and logos, which was used for further analysis and predictive modeling.

C. Data Cleaning

During the process of downloading the logos, some of the retrieved images were not logos but instead depicted unrelated content, such as images of people or natural scenery. These outlier images needed to be removed to ensure that they did not negatively impact analytics or prediction tasks. Given the dataset's scale of approximately 100,000 records, manual cleaning was impractical, necessitating an automated approach with a low false positive rate.

To address this, Centroid Nearest Neighbors was employed to identify and remove outliers in the logo dataset for each industry. The cleaning process involved the following steps:

1) Golden Set Creation: A small set of high-quality, manually verified logo images (approximately 150 per industry) was selected to serve as a "golden set" for each industry. These images acted as representative examples of valid logos for their respective industries.

2) Image Embedding and Similarity Calculation: To compare images, embeddings were extracted using the DenseNet121 [8] model. This model was chosen due to its architecture, which promotes rich feature propagation and reuse by connecting each layer to every other layer. This characteristic made it well-suited for capturing intricate visual features in logo images. Euclidean similarity was then computed between the embeddings of each image in the dataset and those in the golden set.

3) Outlier Identification: Images that were furthest from the centroid of the golden set in terms of cosine similarity were flagged as potential outliers. Approximately 500 images per industry that exhibited the lowest similarity scores were removed from the dataset.

4) Final Dataset: After cleaning, around 9,500 valid logo images remained for each of the 8 industries or business domains, resulting in a refined and reliable dataset. This automated cleaning process ensured that only high-quality logos were retained while minimizing false positives. The cleaned dataset provided a robust foundation for subsequent analysis and predictive modeling tasks.

D. Exploratory Analysis

We analyzed the dataset to uncover interesting relationships between companies within an industry and the colors used in their logos.

Our findings (Fig. 1) revealed that red and blue were the most commonly used colors across all

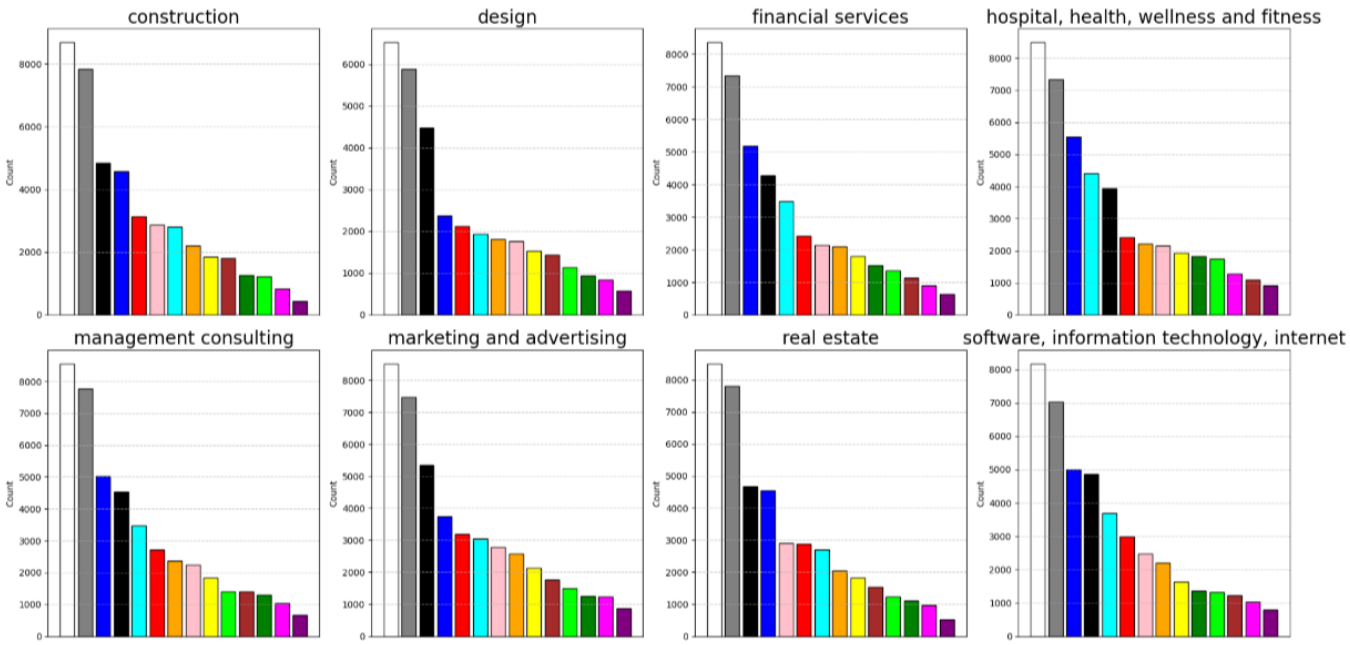

Fig. 1. Colors Used Frequently By Industries

industries, while lime, brown, magenta, and purple were among the least utilized colors in logos across business domains. Red light, having the longest wavelength in the visible spectrum, bends the least compared to other colors like violet. This property results in minimal dispersion for red light, allowing it to maintain its clarity and intensity over greater distances and through various mediums. This characteristic makes red particularly suitable for recognition in both physical and digital media.

The high prevalence of red and blue across industries also indicated that only a limited number of colors are used by most logos across business domains. This suggests that relying solely on color as a feature would not be sufficient to distinctly classify business logos into specific industries.

The color black was used much less in hospital, health, and wellness as compared to other industries. We believe this is because Black is associated with negativity and darkness and industries in the healthcare sector would not be well positioned if their logos were black.

IV. Predicting Business Domain from Logos

A. Baseline: ResNet Embeddings + Logistic Regression

A baseline logistic regression model was trained using ResNet [7] embeddings as features. Since DenseNet embeddings were previously used for outlier removal, we avoided using the same features to prevent any form of data leakage. This ensured that the feature representations used for classification were independent of those used for cleaning the dataset.

The baseline model achieved an accuracy of 24% on the business domain classification task.

The embeddings generated by pre-trained models, such as ResNet, were designed for general image classification tasks and were not specialized for detecting features specific to business logos.

B. Baseline Improvement

Training Specific Classification Models: As a next step, we aimed to enhance the model's capability to detect patterns specific to business logos. To address this problem, we trained a classification model by unfreezing the layers after the embedding layers. This improved the baseline of

24.14% to 24.62%. This improvement could primarily be attributed to the non-linear combination of feature representations achieved through the use of activation functions in neural networks.

C. Result Analysis: Grad-CAM

To understand the behavior of the neural network and analyze what it had learned, we used GradCAM to generate feature activation maps for the trained model on the logos.

As seen in Figure 2, the results revealed that the model was focusing primarily on the whitespace in the logos rather than the design elements or text. This indicated that the model was unable to localize the key features necessary for accurate classification. As a next step, we decided to explore data augmentation techniques to improve feature localization.

D. Whitespace Trimming

The simplest approach to reduce the model's focus on whitespace was to trim the whitespace from the borders of the images before classification. However, after applying this method, the accuracy dropped to 21%. This demonstrated that removing spurious features can sometimes negatively impact accuracy. Additionally, simple feature localization techniques like whitespace trimming did not improve prediction quality, prompting us to explore more sophisticated augmentation methods.

E. Advanced Image Augmentation: CutMix and Mixup

We explored advanced augmentation techniques, specifically Mixup and CutMix, to improve feature localization.

Mixup: Mixup [3] involves creating augmented images by performing a linear interpolation between two images and their corresponding labels. After training the model with Mixup-augmented images, the prediction accuracy improved to 25.26%. Mixup combines images in the training data and this helps to detect distinguishing features of each component image during the training of the model. As noted in prior research, Mixup often produces unnatural images, and the resulting superposition did not appear realistic, hence we also decided to try another augmentation approach.

CutMix: To address this limitation, we explored CutMix [4], an augmentation technique that removes patches from an image and replaces them with patches from another image. The labels are mixed proportionally based on the ratio of pixels used from each image. This method improves feature localization by encouraging the model to focus on partial views of images. After this enhancement, the model's accuracy further improved to 25.67%. Since this is not a superimposition where logos are embossed on each other, the model was able to focus on partial patches containing another logo in a much better way. This could be one of the possible reasons for the slight improvement we are seeing.

When visualizing the feature maps (as seen in figure 3) of CutMix-augmented images revealed stronger activations around text and design elements within the logos. This suggested that CutMix and Mixup augmentations helped the model better localize relevant features for classification.

F. Business Domain Prediction Using Text Extracted from Logos

Upon manually inspecting a few images, we observed that many logos contained text closely related to their respective industries. This observation led to the idea of extracting text from logos and using it for business domain classification.

We utilized the EasyOCR library, which employs CRAFT [2] models for text detection and a Convolutional Recurrent Neural Network (CRNN) for text recognition, to extract the text from the logos. The extracted text was then processed through a DistilBERT [9] encoder model to generate embeddings. These embeddings were subsequently used to train classifiers for predicting the industry of each logo.

This approach significantly improved accuracy, increasing it from the baseline score of 24% to 41.22%. These results demonstrate that the text within logos provides valuable information about their corresponding business domains, making it a critical feature for classification.

From the Confusion Matrix (Fig 6), it is visible that the model performs best while classifying the domain 'hospital, health, wellness and fitness'. This can be attributed to the fact that many of the logos from this domain consist of terms like

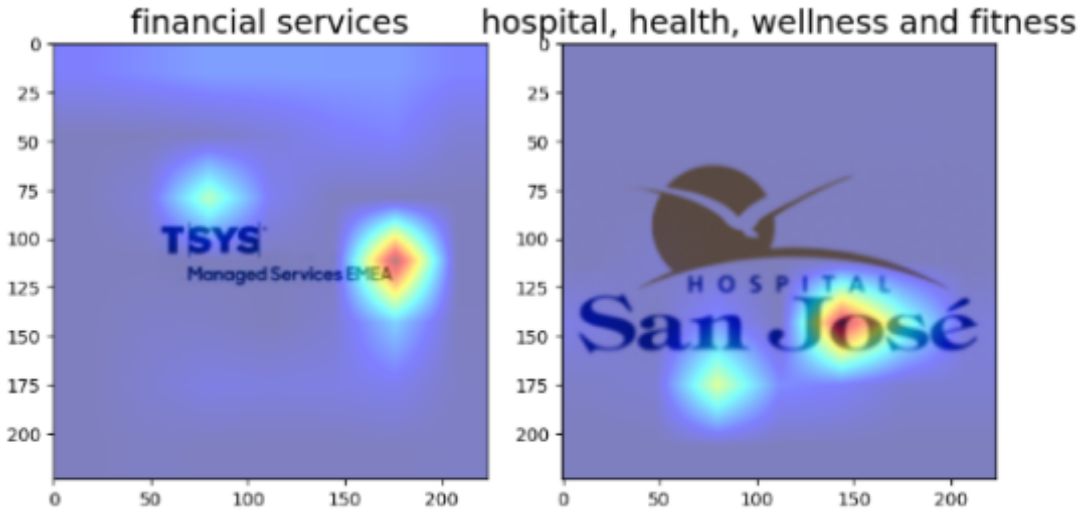

Fig. 2. Grad Cam Without Image Augmentation

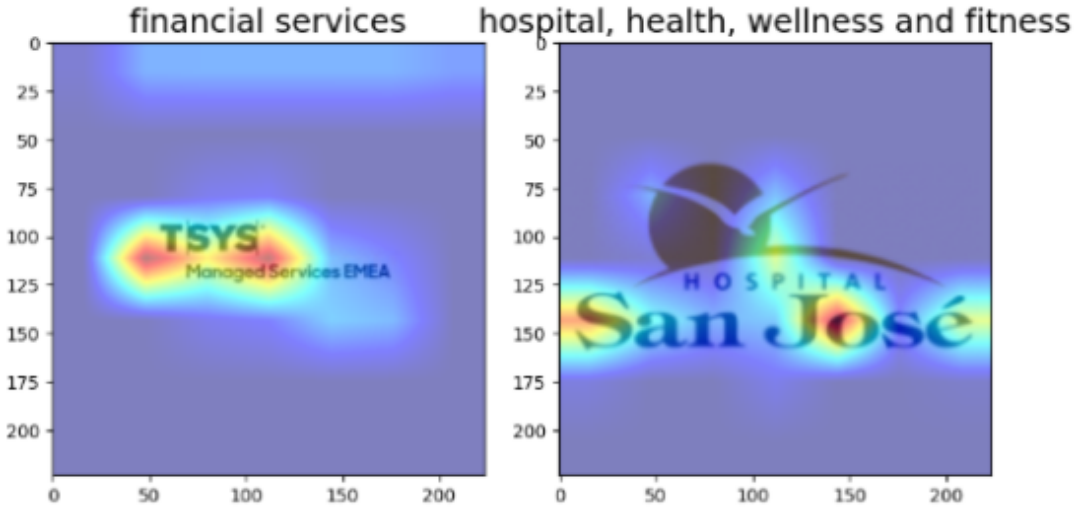

Fig. 3. Grad Cam With Image Augmentation

Fig. 4. MixUp Results for 2 Images

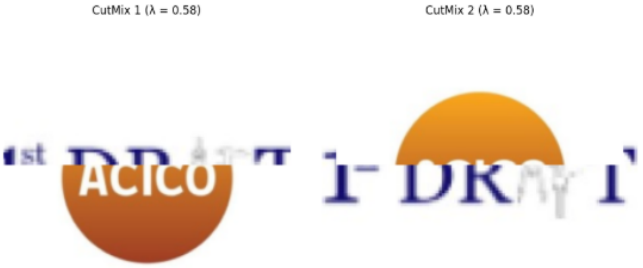

Fig. 5. CutMix Results for 2 Images

'health', 'healthcare', 'hospital', 'medicine', etc. These terms are absent in other classes and therefore the model does a better job classifying for this class. We can also see that the model mistakes the classes 'design' and 'marketing and advertising'. This is because these 2 classes are somewhat similar and hence would use many common words in their logos.

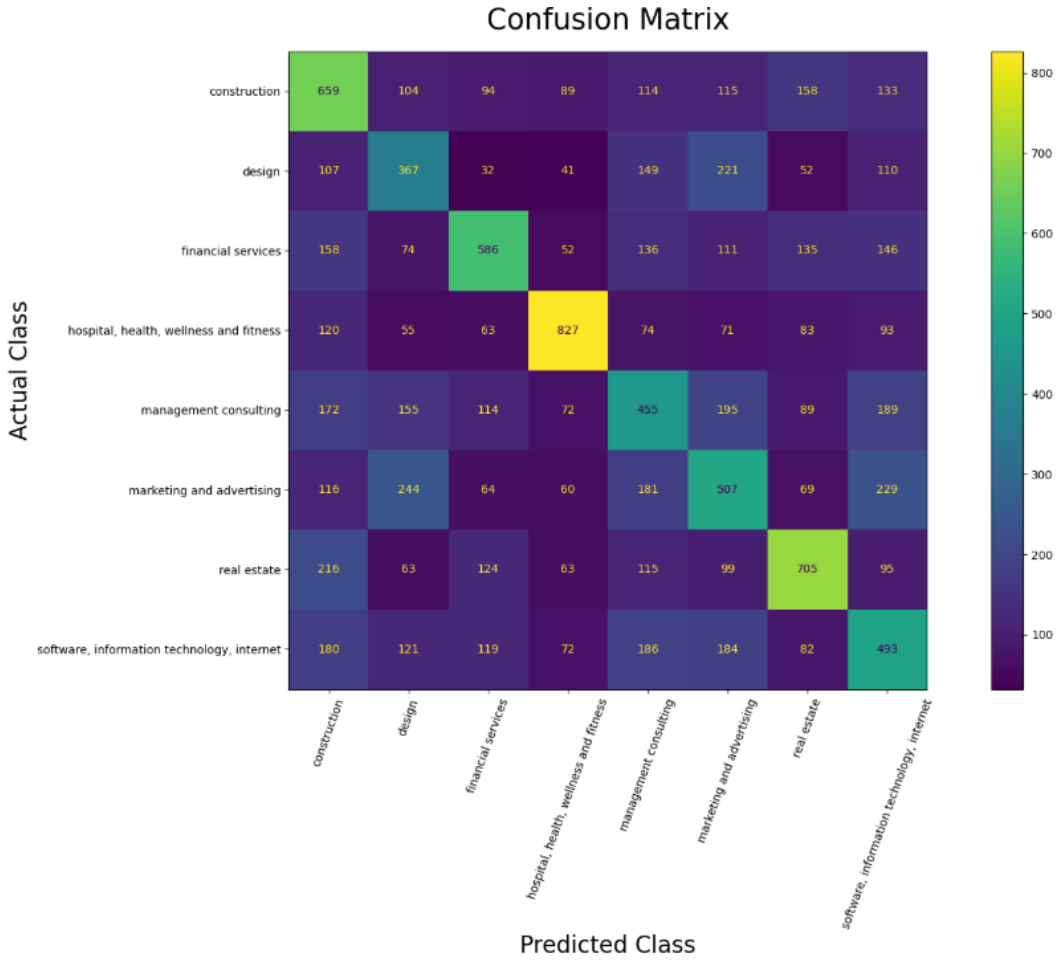

Fig. 6. Confusion Matrix for DistilBERT

TABLE I
SUMMARY OF BUSINESS DOMAIN CLASSIFICATION RESULTS

Experiment Description	**Accuracy (%)**
Logistic Regression with ResNet embeddings	24.14
ResNet50 model fine-tuned on logo dataset	24.62
ResNet50 fine-tuned with Mixup	25.26
ResNet50 fine-tuned with CutMix	25.67
Logo text extraction with OCR + DistilBERT text embeddings	41.22

G. Evaluation of Statistical Significance

We executed a permutation test with 1000 iterations on the predictions and achieved a p-value of 0.0. This p-value is less than the level of significance (0.05), allowing us to reject the null hypothesis and conclude that our results are statistically significant.

V. CONCLUSION

In this project, we created a comprehensive dataset of companies, their industries, and corresponding business logos. Using this dataset, we developed and implemented prediction models to classify business domains based on the features of their logos. Our best-performing model achieved an accuracy of 41.22%, highlighting a clear relationship between the text within business logos and their associated domains. To ensure the reliability of our findings, we also evaluated the statistical significance of the results using a permutation test. This study underscores the importance of textual elements in logos for domain classification while paving the way for further exploration to enhance model accuracy.

VI. Future Work

We used pretrained models like ResNet and DenseNet, which were originally trained on general image datasets such as ImageNet. These models may not be optimized for logo-specific features.

Incorporating a specialized model like DRNA-Net (Discriminative Region Navigation and Augmentation Network) [6], which is designed to focus on informative logo regions and augment them for better feature extraction, could significantly enhance logo classification performance. DRNA-Net's ability to localize and process logo-relevant regions makes it well-suited for this task.

Additionally, exploring multimodal approaches by combining visual features from logos with textual information extracted from them could further improve classification accuracy. For instance, using Vision-and-Language Transformers (ViLT) [10] to process both image embeddings and text embeddings simultaneously could enable fine-grained interactions between modalities, leading to better predictions.

Another promising direction is analyzing font styles within logos. Fonts often convey industry-specific characteristics (e.g., playful fonts for toy companies or formal serif fonts for newspapers). Developing a font extraction pipeline could provide additional features to improve classification performance.

VII. Code

The code for the Project is uploaded at: https://github.com/Vignesh1399/Business-Logos/.

References

[1] Khani, Fereshte, and Percy Liang. "Removing spurious features can hurt accuracy and affect groups disproportionately." In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 196-205. 2021.

[2] Baek, Youngmin, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. "Character region awareness for text detection." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9365-9374. 2019.

[3] Zhang, Hongyi. "mixup: Beyond empirical risk minimization." arXiv preprint arXiv:1710.09412 (2017).

[4] Yun, Sangdoo, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. "Cutmix: Regularization strategy to train strong classifiers with localizable features." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 6023-6032. 2019.

[5] Shi, Baoguang, Xiang Bai, and Cong Yao. "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition." IEEE transactions on pattern analysis and machine intelligence 39, no. 11 (2016): 2298-2304.

[6] Wang, Jing, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, Haishuai Wang, and Shuqiang Jiang. "Logo-2k+: A large-scale logo dataset for scalable logo classification." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 04, pp. 6194-6201. 2020.

[7] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[8] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely connected convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708. 2017.

[9] Sanh, V. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).

[10] Kim, Wonjae, Bokyung Son, and Ildoo Kim. "Vilt: Vision-and-language transformer without convolution or region supervision." In International conference on machine learning, pp. 5583-5594. PMLR, 2021.

[11] https://www.kaggle.com/datasets/peopledatalabssf/free-7-million-company-dataset

[12] Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: visual explanations from deep networks via gradient-based localization." International journal of computer vision 128 (2020): 336-359.

[13] https://github.com/JaidedAI/EasyOCR

[14] Mohd Ali, Nursabillilah. Performance Comparison between RGB and HSV Color Segmentations for Road Signs Detection. Applied Mechanics and Materials. 393. 10.4028/www.scientific.net/AMM.393.550. (2013).

[15] A. Ajmal, C. Hollitt, M. Frean and H. Al-Sahaf, "A Comparison of RGB and HSV Colour Spaces for Visual Attention Models," 2018 International Conference on Image and Vision Computing New Zealand (IVCNZ), Auckland, New Zealand, 2018, pp. 1-6, doi: 10.1109/IVCNZ.2018.8634752.

[16] Park, C. Whan, Andreas B. Eisingerich, Gratiana Pol, and Jason Whan Park. "The role of brand logos in firm performance." Journal of business research 66, no. 2 (2013): 180-187.

[17] ÇELİKKOL, Şimal. "The importance of logos and strategies for logo design." POLITICO-ECONOMIC EVALUATION OF CURRENT ISSUES (2018): 29.

[18] Vedhaviyassh, D. R., R. Sudhan, G. Saranya, Mozhgan Safa, and D. Arun. "Comparative analysis of easyocr and tesseractocr for automatic license plate recognition using deep learning algorithm." In 2022 6th International Conference on Electronics, Communication and Aerospace Technology, pp. 966-971. IEEE, 2022.

[19] Smith, Ray. "An overview of the Tesseract OCR engine." In Ninth international conference on document analysis and recognition (ICDAR 2007), vol. 2, pp. 629-633. IEEE, 2007.

[20] Hou, Sujuan, Jiacheng Li, Weiqing Min, Qiang Hou, Yanna Zhao, Yuanjie Zheng, and Shuqiang Jiang. "Deep learning for logo detection: A survey." ACM Transactions on Multimedia Computing, Communications and Applications 20, no. 3 (2023): 1-23.