IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 1, 2022, accepted January 14, 2022, date of publication January 18, 2022, date of current version February 1, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3144185

A Multi-Modal Approach to Digital Document Stream Segmentation for Title Insurance Domain

ABHIJIT GUHA[1,2], **ABDULRAHMAN ALAHMADI**[3], **DEBABRATA SAMANTA**[4], **(Member, IEEE)**, **MOHAMMAD ZUBAIR KHAN**[3], **AND AHMED H. ALAHMADI**[3]

[1]Department of Data Science, CHRIST (Deemed to be University), Bengaluru, Karnataka 560029, India
[2]First American India Private Ltd., Bengaluru, Karnataka 560038, India
[3]Department of Computer Science and Information, Taibah University, Medina 42353, Saudi Arabia
[4]Department of Computer Science, CHRIST (Deemed to be University), Bengaluru, Karnataka 560029, India

Corresponding authors: Debabrata Samanta (debabrata.samanta369@gmail.com) and Mohammad Zubair Khan (mkhanb@taibahu.edu.sa)

ABSTRACT In the twenty-first century, storing and managing digital documents has become commonplace for all corporate and public sectors around the world. Physical documents are scanned in batches and stored in a digital archive as a heterogeneous document stream, referred to as a digital package. To make Robotic Process Automation (RPA) easier, it's necessary to automatically segment the document stream into a subset of independent, coherent multi-page documents by detecting the appropriate document boundary. It's a common requirement of a TI company's Automated Document Management Systems (ADMS), where business operations are automated using RPA and the goal is to extract information from digital documents with minimal user intervention. The current study proposes, evaluates, and compares a multi-modal binary classification network incorporating text and picture aspects of digital document pages to state-of-the-art baseline methodologies. Image and textual features are extracted simultaneously from the input document image by passing them through Visual Geometry Group 16 - Convolutional Neural Network (VGG16-CNN) and pre-trained Bidirectional Encoder Representations from Transformers (Legal-BERT$_{base}$) model through transfer learning respectively. Both features are finally fused and passed through a fully connected layer of Multi Layered Perceptron (MLP) to obtain the binary classification of the pages as the First Page (FP) and Other Page (OP). Real-time document image streams from production business process archive were obtained from a reputed Title Insurance (TI) company for the study. The obtained F_1 score of 97.37% and 97.15% are significantly higher than the accuracies of the considered two baseline models and well above the expected Straight Through Pass (STP) threshold defined by the process admin.

INDEX TERMS Page stream segmentation, multi modal training, binary classification, title insurance, BERT, VGG16.

I. INTRODUCTION

Despite the colossal growth of machine intelligence in the recent past, many tasks seem too abstruse when handled by machines but performed effortlessly by human beings. Document Stream Segmentation (DSS) is one of them. Multi-page digital documents arrive at the Document Management System (DMS) as an ordered set of digital images without any indication of the document boundaries. DSS is the task of breaking the page stream into a set of documents. Traditional DMSs needed human intervention to place the page-break indicators (extra page or bar codes) at the source of digitization for machines to be able to perform the segmentation during the real-time processing [1]. This is a costly affair and not feasible for all digitization sources. In the wake of AI, with RPA integration, ADMSs are fast replacing DMSs [2]. ADMS is sometimes referred to as Intelligent Document Processing Systems (IDPS). Like other business domains, ADMS is an important component of the TI. Examination of digitized document packages comprised of multiple documents of varying length and quality is ubiquitous in a typical TI search and examination process. There is a need for segmenting the packages devoid of any preset indicators automatically into individual documents so that the subsequent

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

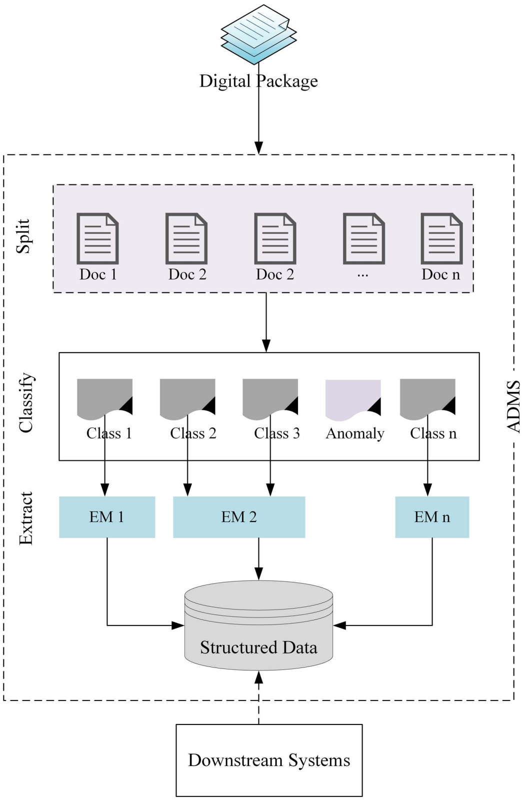


FIGURE 1. Typical flow of a package within ADMS.

modules (Classification, Extraction) of the ADMS can act for additional automated processing (Figure1). An efficacious DSS is a need for any ADMS to be precise in the consecutive tasks because any error occurring in the DSS has a rippling effect on the subsequent modules bringing the overall precision of the system down.

We have traced the evolution (Table 1) of the DSS technologies starting from the stochastic Markov chain model in 2009 [3] through the deep image-based page feature extraction and classification in 2016 [4], rule-based approach in 2017 [5] to a more sophisticated state-of-the-art multi-modal deep learning approach combining text and image features of the document page until 2021 [6]. Although, there is a clear dearth of the overall study observed in the domain of DSS, the recent breakthrough in the domain of DSS by Wiedemann and Heyer [7] shows promising results with Tobacco800 public data set and a proprietary data set. The work of Braz *et al.* is built over the proposal of Wiedemann and Heyer [7] by improving the network architecture by using EfficientNet pre-trained CNN architecture, replacing the earlier proposed VGG network. We evaluated our proprietary TI data set with both the proposed architectures and received maximum accuracy of 86.7% which is significantly lower than the expected accuracy. Having analyzed the result, we observe certain research gaps and found the further scope of study which motivated the proposed work.

There are not many publicly available data set for DSS in the legal and property domain. The proposed architectures did not perform at the same level of accuracy with our data set as claimed by the past researchers. With a real-time problem in hands from a reputed TI company with its proprietary data, it was important for us to explore the problem further. It has been empirically established that the textual features have more prominence towards document image classification contrary to image classification [9], [10]. The image features improve the classification results in a multi-modal environment but cannot alone perform the job [11], [12]. However, the past researchers in the multi-modal context exerted themselves into various image networks keeping the text feature extractor almost unexplored even though the state-of-the-art in the Natural Language Processing (NLP) technology domain has seen a paradigm shift with many transformers-based models like Bidirectional Encoder Representations from Transformers (BERT). The past researchers adopted MLP and Support Vector Machine (SVM) for the binary classification of the multi-modal page feature. The MLP architectures have been kept simple and single-layered. There is a scope to experiment with the final MLP layer with a variation of depth and size of the hidden layers. DSS is a common task in ADMS of the Title Insurance industry as the documents are stored and shared within the business divisions in bulk streams. It is tedious and costly for the process associates to segment the documents from the heterogeneous page stream. Motivated by the real-time need for the business processes and problems of the associates of the TI industry, the research is performed.

In this work, we have considered the model proposed by Wiedemann and Heyer [7] as the baseline. Multiple models are trained and evaluated keeping the image feature extractor as a VGG16 pre-trained model. However, the text feature is replaced by the pre-trained Legal-BERT$_{base}$ [8] model. The segmentation algorithm depends on the page level output of the page class as FP/ OP. Our proposed model took the advantage of state-of-the-art transfer learning from both modes of feature representation.

A unique, real-time proprietary data set used in the TI closing and examination business process with the mortgage, legal, and property title-related documents have been evaluated. Due to legal bindings, the data can not be made public. However, the proposed technique can be evaluated by the research community. To our knowledge, no prior DSS study has been conducted to explore and validate the data in Title Insurance domain. The proposed model has not only empirically established the superiority of the DSS task but also is deployed in the production environment of the organization in the IDPS setup. Transformer technology, specifically BERT is the latest state-of-the-art for text feature embedding in the NLP domain. This has not been adopted in the DSS study in the past. Adopting transfer learning using BERT has been one of the key novelty of the proposed study.

The remainder of this paper is organized into six subsequent sections. In Section 2, background study of Page

TABLE 1. Evolution of DSS research with milestone works since 2009.

Publication	Year	Techniques/ Features	Data set	Reported Efficacy
Braz et al.	2021	Textual and image feature with EfficientNet-B0	AI.Lab.Splitter data set from Brazilian Judiciary system	93.5 % - 95.5% (F_1 score)
Wiedemann et al.	2017	Bi-modal text and image feature with CNN	German Federal Archive and Tobacco800 US	91% and 93% (Accuracy)
Karpinski et al.	2016	Textual, physical, logical and factual descriptors	Corpus from ITESOFT.	$75\% - 96.17\%$ (precision)
Gallo et al.	2016	CNN+ DNN	Public Dataset	88.16% (Accuracy)
Agin et al.	2015	BoVW feature descriptor with SVM, RDF and MLP	Banking data set.- Private DS	88.88% (F_1 score)
Rusiñol et al.	2014	Text + Image with SVM	National Institute of Standards and Technology (NIST) Tax form (SPDB2), MARG and In-house private dataset	84.26% - 96.84% (Accuracy)
Daher et al.	2014	Textual descriptors	Four databases of ITESOFT.-Private DS	84% (homogeneity index)
Gordo et al.	2013	Text+ Image	Administrative documents from banking domain.- Private DS	22.35 (mAP)
Meilender et al.	2009	Markov Chain Model	Homogeneous stream of invoices from ITESOFT.- Private DS	75% (precision)
Proposed work	2021	Image (VGG16) + Text (Legal-BERT$_{base}$ [8]) + MLP	Private DS from TI archive CP2020, SP2020	97.37%, 97.15% (F_1 score)

stream segmentation is articulated in an evolutionary format, followed by the problem statement and materials and methods adopted with the architecture for the current research in section 3 and 4. A detailed explanation of the experimental setup and experiments conducted with different models are provided in section 5, followed by dissecting the results in Section 6 and conclusive remarks with future scope in Section 7.

II. RELATED WORKS

Various experimental studies in the related domain have been conducted in the recent past, popularly known as DSS or Document Flow Segmentation (DFS). The experiment techniques adopted in those studies can be broadly categorized into two groups;Rule-based systems, and Machine Learning-based systems. Three prominent procedures are observed under machine learning-based techniques. Some studies depended on textual features and some on image features.

Wiedemann *et al.* took a hybrid approach of combining image and the text-based features with building a single classification architecture [13] to perform the task of DSS. The system was tested with both in-house and public data sets, and the authors achieved an accuracy of 95% and 93%, respectively. Text-based SVM based classifier was considered the baseline in the study and was compared with CNN for multi-modal DSS with the combined features.

A contextual and layout descriptor-based approach that represented the relationship of two consecutive pages of document stream was presented by Hamdi *et al.* [5], [14]. In this approach, every page was represented with binary features of contextual and layout information, such as the textual fingerprint, ending signs, page number, dates, etc. A two-class classifier was trained using a decision tree to classify the pages into either a continuation or a break class where continuation class determines a page to be a continuation of the previous page, and break class determines the beginning of a new document. In a continuous effort to find the best approach, the authors compared the segmentation result using both rule-based and a machine learning-based approach to define the features and found the machine learning-based approach to produce better results than the rule-based approach [5].

A similar approach of comparing the pages based on textual, physical, logical, and factual descriptors is adopted by Karpinski *et al.* [15]. Additionally, if a page suffers from information emptiness, the authors proposed a lazy comparison of the page similarities by maintaining a logbook approach to keeping the previous page information.

Gallo *et al.* proposed in their study a hybrid technique using CNN, followed by a Deep Neural Network (DNN) to extract the visual features of the text and classified the documents. The page stream segmentation was done using the document classification in the proposed method [4].

Bag of Visual Words (BoVW) combined with the font features was used as descriptors for every document page to classify the page as a new page or continuation of the previous page has been studied by Agin *et al.* SVM, Random Forest (RBF) and MLP are evaluated for the binary classification task. The random forest has achieved the highest F_1 score as captured in the experimental results [16]. A very similar study has been carried out by Daher *et al.* [17], [18]. The authors finalized nine feature descriptors for the pages and used a regular expression to extract those. The pages represented in the nine-dimensional feature space were trained with voted perceptron, SVM, MLP, and multi-boost algorithms, and the results were compared.

Rusinol *et al.* presented their study on a real-time application in banking workflow [19]. The authors proposed an architecture that combined the visual and the text feature for the feature descriptor. The visual part of the representation was based on the hierarchical pixel intensity distribution, and the textual part was based on latent semantic analysis for topic understanding. The result was evaluated upon experimenting with a large real-time data set of 70,000 pages. In the first stage of the model, the visual and text features are independently predicted. The prediction probabilities are combined with the n-gram features in the latter stage of the model to classify the document.

Document separation from a stream of batch scanned documents is the task of a digital mailroom system. But the automation comes with the cost of tagging the separation boundary during the scanning process. Gordo *et al.* [1] thus proposed a supervised approach of classifying document pages into appropriate class and associated the solution to the document separation from the digital stream of pages. The classification was experimented with both textual and image features of the document pages.

A novel approach of document stream segmentation using a Variable Horizon Model (VHM) [3], popular in speech recognition, was proposed by Meilender *et al.* The approach maximized the flow likelihood using Markov models and, based on the likelihood, separated the page streams into individual documents. A very similar but hybrid technique was proposed by Schmidtler *et al.* The authors proposed a combination of probabilistic classification and sequence modeling to achieve the document separation task [20].

The authors accept that a generic system for document segmentation from a stream of scanned pages is a challenging task [19]. The rule-based systems tend to overfit based on the type of documents the solution deals with. In a highly heterogeneous domain like TI where there are various documents or varying lengths, the generic solutions are producing results that are not up to the mark. Also, it is a very tedious task to define a generic set of features that define the pages of various documents. We observed that there are a limited set of documents that every business processes use over time in TI. For example, an insurance closing process deals with approximately one hundred twenty to one hundred fifty types of documents, whereas a search process deals with about ninety types of documents. In this study, we propose a simple supervised approach that is more effective on a single process level.

III. MATERIALS AND METHODS

A. THE PROBLEM

The DSS problem relates to identifying a page from a stream of pages as either the start of a new document or the continuation of the previous document. A human associate without any prior domain knowledge, just by looking at certain visual and textual features should be able to determine the page as either of the above two classes. The algorithm of DSS also adopts the same philosophy.

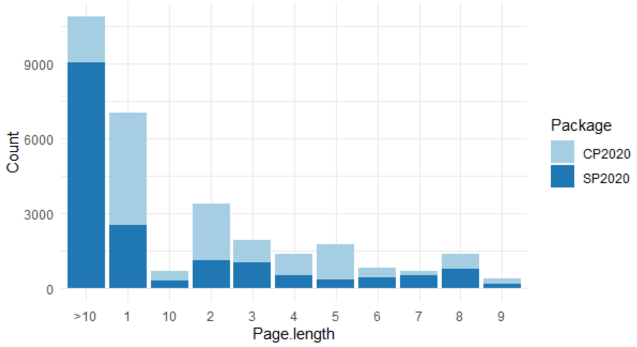


FIGURE 2. Page length distribution for the documents from Archive-CP2020 and Archive-SP2020.

Formally, in literature, DSS is defined as a function $\mathbb{F} : \mathbb{P} \rightarrow \mathbb{D}$, where $\mathbb{P} = \{p_1, p_2, \ldots, p_N\}$ is a set of N pages transformed to $\mathbb{D} = \{d_1, d_2, \ldots, d_M\}$, a set of M multi-page documents of sequential pages, using a binary classification function $\mathbb{G}(X, Y)$, where $d_k = \{p_i, p_{i+1}, \ldots, p_j\}$ for $i \leq j$ and $X \in \mathbb{R}^n$, $Y \in \{0, 1\}$. Here, 1 denotes the first page of any document and 0 denotes any page other than the first page of any document [4], [6].

Although, the past studies on DSS considered the multi-page document classification post segmentation also part of the algorithm, our proposal is limited to only segmentation of the documents into cohesive multi-page subsets. We consider classification of multi-page documents an independent and already matured research domain.

B. DATA SET

Since the study is motivated by a real-time problem of a reputed TI company wherein their business process DSS is a typical problem that needs immediate attention to facilitate and improve the RPA, we took the real-time data from the business process archive of the same company for the study. Firstly, we evaluated the baseline with the data set, and after having found inadequacy in the accuracy, we proposed the alternate architecture for the DSS. Two sets of packages were considered from the archive of 2020 January to June. CP2020 is the package consisting of document streams arriving at the closing business process whereas SP2020 is that of the search business process. Statistics of both the samples are explained in the table (Table 2), and the distribution of the length-wise document count is captured in the Figure (Figure2).

Closing Packages arrive at the business divisions from the lenders during the closure of an insurance policy. Different types of documents related to loan closure, for example, closing instructions, closing disclosure, HUD documents, settlement statement, uniform residential loan applications, etc. are the document types generally associated with the package. In CP2020, we have encountered 139 different document types associated across all packages within the samples. FP and OP image of some example documents are shown in figure 3 and 4.

TABLE 2. Quantitative description of the data sets.

Archive-CP2020	
Documents	13, 579
Single Page	4481(32.99%)
Multi Page	9098(67.01%)
Page length of documents	**Page Count**
1	4481
2	4564
3	2745
4	3328
5	7000
6	2268
7	1421
8	4928
9	1782
10	4000
> 10	43, 222
Archive-SP2020	
Documents	16, 858
Single Page	2546(15.10%)
Multi Page	14, 312(84.89%)
Page length of documents	**Page Count**
1	2546
2	2244
3	3105
4	2176
5	1850
6	2592
7	3577
8	6264
9	1701
10	3000
> 10	76, 678
Total Documents:	30, 437
Total Pages:	185, 472

Search packages are used in the title examination process during the TI order creation. As soon as an order is created in a TI order entry system, various historical documents associated with the said property are pulled from different private and public document sources and archives. Past deeds, mortgages, releases, tax receipts, liens, judgements etc. are some common document types present in this package. 187 document types are found in SP2020 samples. Some examples of FP and OP of the documents are shown in figure 5 and figure 6.

The experiment is conducted in a four-step method; Data collection, ground truth generation, training, and testing. CP2020 and SP2020 data sets are collected and manually annotated as FP and OP with 80% and 20% train and test split. The training data is then trained with the proposed multi-modal architecture followed by binary classification and validation. The text and image features are fused before the binary classification layer and trained within a combined feature space. Finally, the prediction by the model for a document image page is either FP or OP, using which the page stream splitting is carried out as described in the algorithm 1.

C. ARCHITECTURE

Human intelligence is multi-modal. We perceive the world through different modalities such as hearing, tasting, smelling, touching, and seeing. Therefore, real-world problems that are being replaced by artificial intelligence also need to be multi-modal in nature. DSS is one such task that is performed by human associates by looking at the texts of the documents as well as the visual features of the page. Usually, the starting page of any document holds distinctive visual and textual features. Logos, bigger title fonts, page numbers help us detecting the page boundary pretty effortlessly. Sometimes, these differentiating features are domain-specific [2]. The general approach of implementing such a multi-modal solution is to concatenate the signal embeddings and pass the hybrid features through Softmax or Sigmoid function for multi-class or binary classification.

Mathematically, M denotes the number of modalities. Each modality is represented by dense vector $v_m \in \mathbb{R}^{d_m}$, $\forall m = 1, 2, \ldots, M$. In this study, $M = 2$, and v_1, v_2 represents the feature vectors from text and image of the document pages respectively. For a k class classification scenario p_m^k represents the probability of k^{th} class for modality m and p^k denotes the overall probability of k^{th} class denoted by label y. $K = 2$ represents a binary classification in the current proposal. Traditionally, the aforesaid multi-modal training is performed in two ways; Early fusion and Late fusion. In early fusion scenario, a model h is trained on a joint representation of features from m modalities (equation 1) whereas, in the late fusion method, independent models h_m are trained for m modalities and the decisions are fused through techniques like averaging, voting or further training represented by a function κ (equation 2) [21], [22].

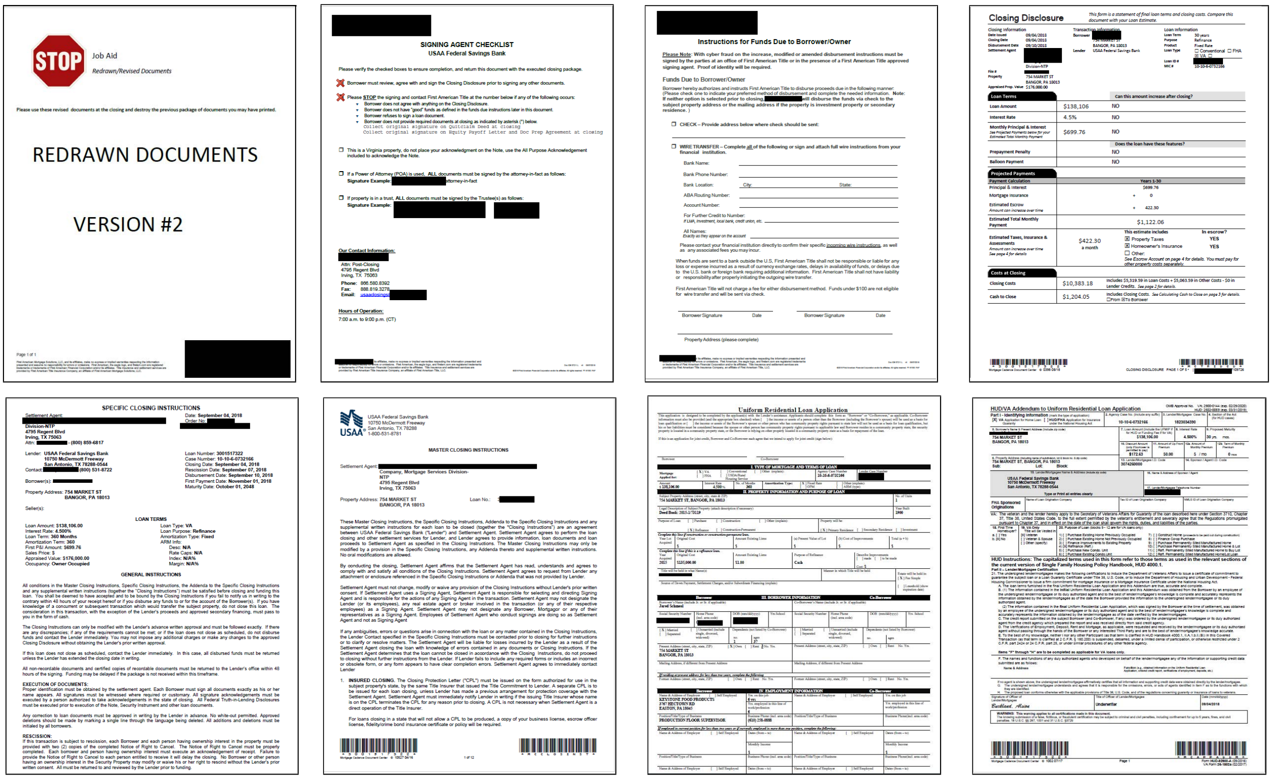
$$p = h([v_1, \ldots, v_m]) \tag{1}$$

$$p = \kappa(h_1(v_1), \ldots, h_m(v_m)) \tag{2}$$

1) IMAGE FEATURE EXTRACTOR USING CNN (VGG16)

We followed the same modality for image data proposed by Wiedemann and Heyer *et al.* They put forward the VGG16 CNN network for the image data extraction [13]. Karen Simonyan and Andrew Zisserman of the VGG lab presented this architecture in ILSVRC 2014. The model reported 92.7% accuracy in image classification, and object detection task on 14 billion images of ImageNet data set [23]. Wiedemann *et al.*, in their work, have applied Otsu's binarization as a pre-processing step before sending the image into the VGG16 network. However, we retained the RGB color channels of the input image. The color features of various logos, predominantly present on the first page of the documents could be a distinct discriminating factor for the images.We have replaced the binarization step with skew correction to standardize the orientation of the documents.

The input layer dimension of VGG16 is 224×224 followed by two Conv. layers of 128 channels with filter size 3 × 3. A 2 × 2 max-pool layer with stride 2 is inserted after this.


Job Aid

Redrawn/Revised Documents

REDRAWN DOCUMENTS

VERSION #2

SIGNING AGENT CHECKLIST
USAA Federal Savings Bank

Instructions for Funds Due to Borrower/Owner

Closing Disclosure

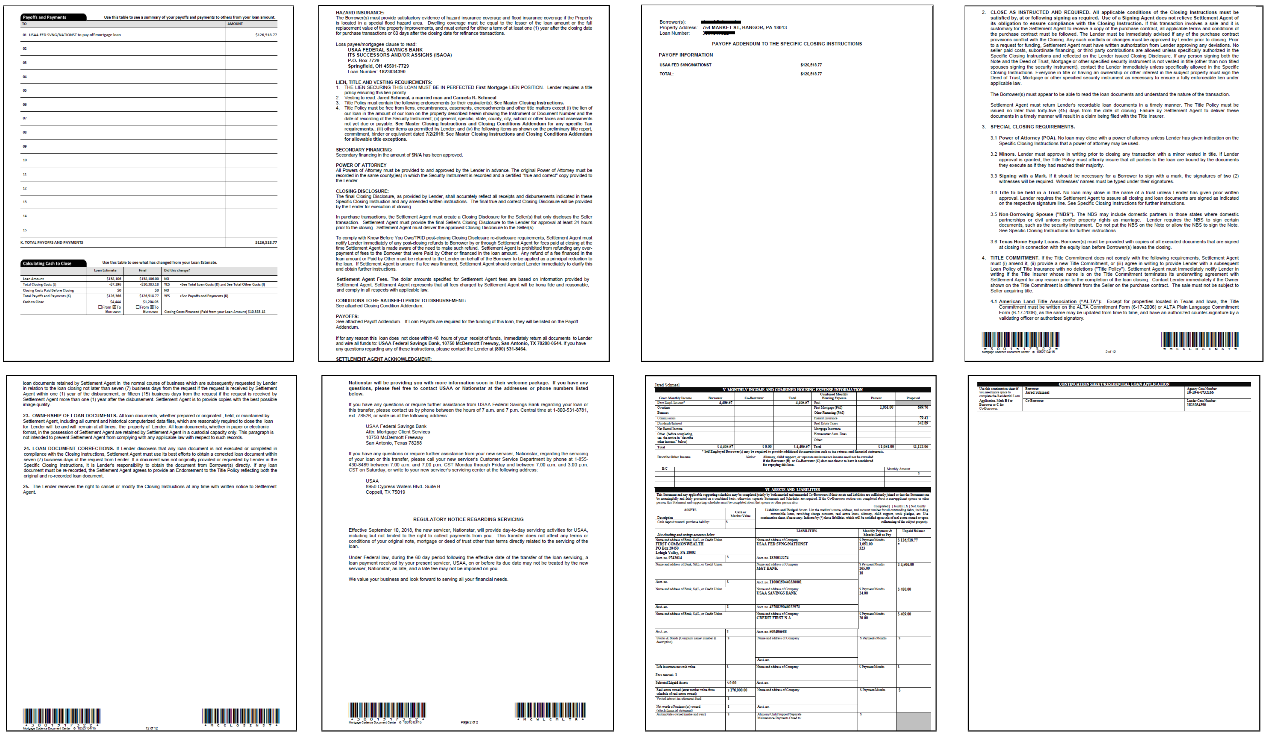
SPECIFIC CLOSING INSTRUCTIONS

MASTER CLOSING INSTRUCTIONS

Uniform Residential Loan Application

HUD/VA Addendum to Uniform Residential Loan Application

FIGURE 3. Example images of first pages of job aid, agent checklist, instructions to borrowers, closing disclosure, specific closing instructions, master closing instructions, uniform residential loan application and addendum to uniform residential loan application from Archive-CP2020. The Non Public Information (NPI) and Personal Information (PI) information are redacted for info-sec purposes. Being the first page of the documents, the presence of logo and difference in font size is noticed. These visual features are expected to be learnt by the model.


PAYOFF ADDENDUM TO THE SPECIFIC CLOSING INSTRUCTIONS

REGULATORY NOTICE REGARDING SERVICING

FIGURE 4. Example images of other pages of agent checklist, instructions to borrowers, closing disclosure, specific closing instructions, master closing instructions, uniform residential loan application and addendum to uniform residential loan application from Archive-CP2020. The NPI and PI information are redacted for info-sec purposes. Being the non first page of the documents, the absence of logo and similar font size is noticed. These visual features are expected to be learnt by the model.

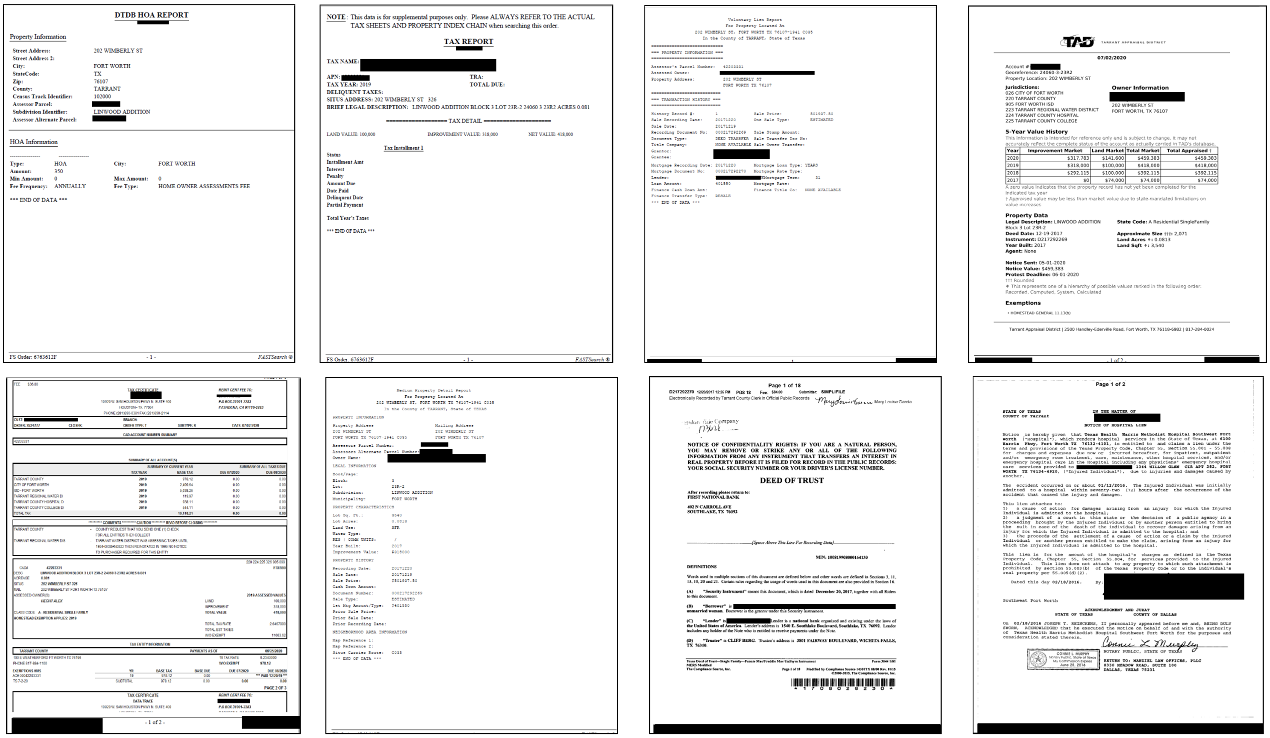


FIGURE 5. Example images of first pages of tax report, voluntary lien report, medium property detail report, deed of trust, property index report, lien notice from Archive-SP2020. The NPI and PI information are redacted for info-sec purposes.Being the first page of the documents, the presence of logo and difference in font size is noticed. These visual features are expected to be learnt by the model.

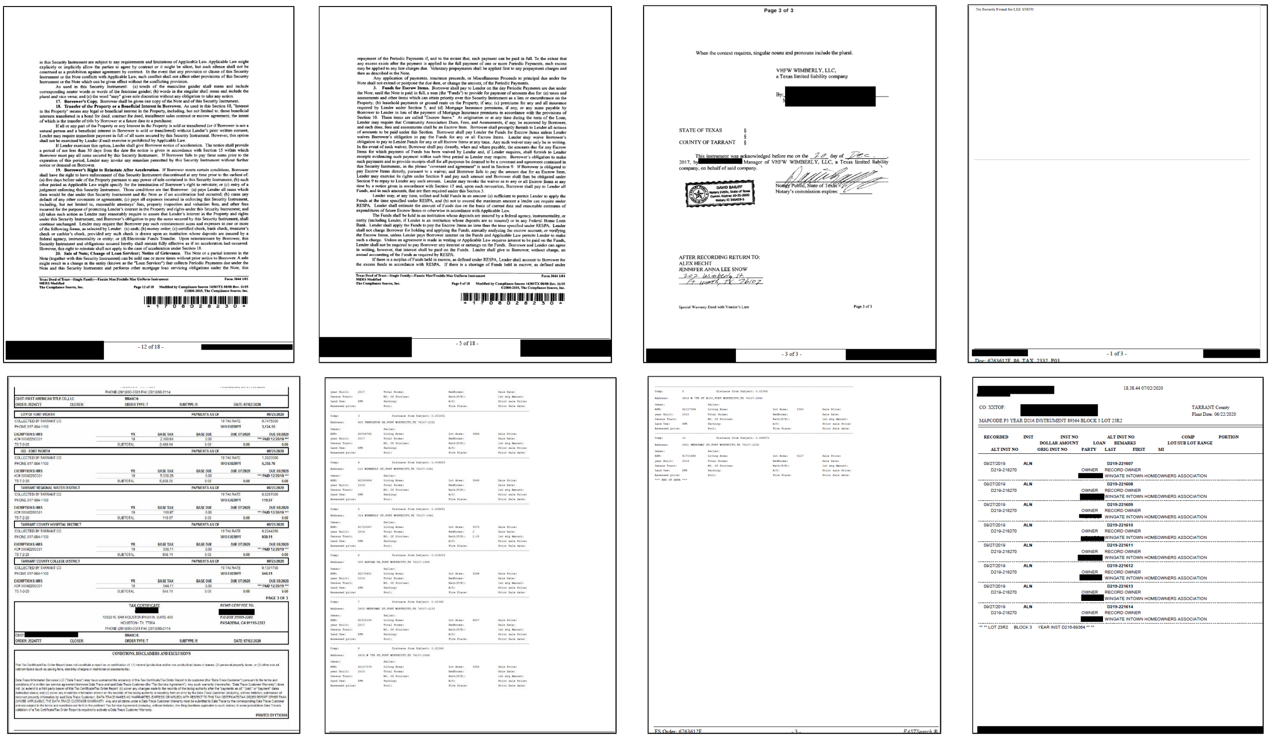


FIGURE 6. Example images of other pages of tax report, voluntary lien report, medium property detail report, deed of trust, property index report, lien notice from Archive-SP2020. The NPI and PI information are redacted for info-sec purposes.Being the non first page of the documents, the absence of logo and similar font size is noticed. These visual features are expected to be learnt by the model.

The same max-pool configuration is used in all the five layers of VGG16 as shown in the Figure 7. Rectified Linear Activation Unit (ReLU) activation is used throughout the hidden layers in the proposed architecture.In the final layer, two Fully-Connected dense layers with 4096 nodes each and one layer with 1000 dense nodes are placed. In the original

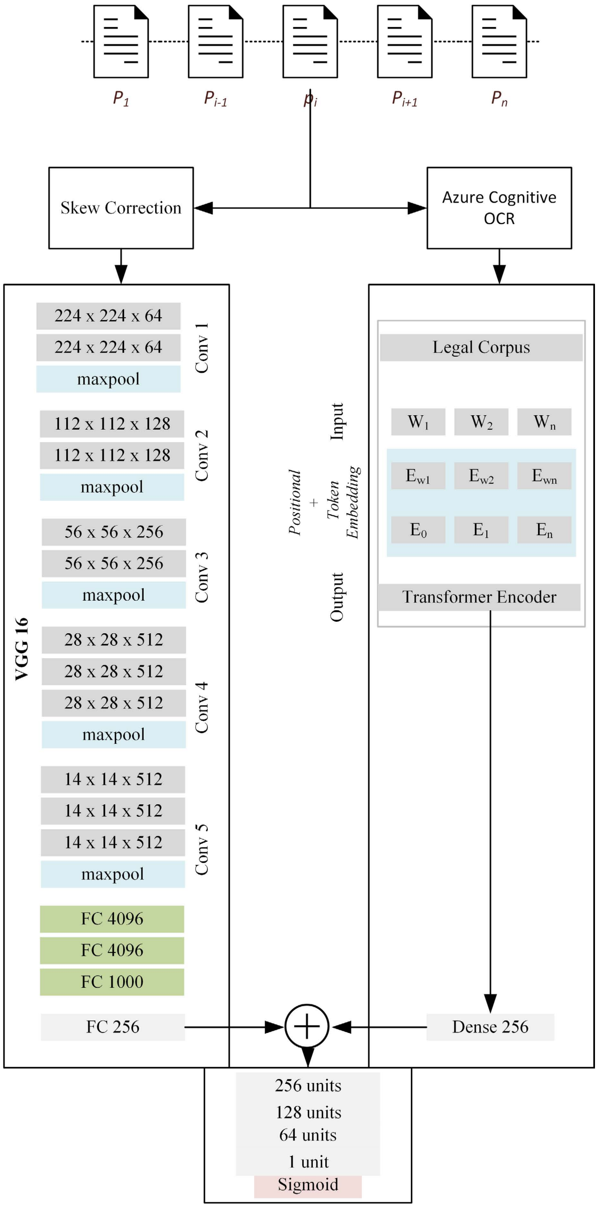


FIGURE 7. Multi-modal architecture of DSS with pre-trained VGG16 and Legal-BERT$_{base}$.

architecture, a soft-max layer is incorporated to classify the images in 1000 classes. However, we have disconnected that classification layer as we need the embedding to be passed to a dedicated MLP for the binary classification task in the subsequent phase of the model. Although, VGG16 has not been optimized for classifying the document images our study empirically shows that it helps to improve the image document classification with its pre-trained weights. A trainable layer of 256 units is added on top of the above network with 0.3 dropout regularization. The network has 138 million parameters to train [24], [25].

2) TEXT FEATURE EXTRACTOR (LEGAL-BERT$_{base}$) [8]

BERT is the new state-of-the-art technology based on 'Attention' [26] and 'Transformers' [27] for various NLP and Natural Language Understanding (NLU) tasks. It is intended to train deep bidirectional illustrations from unsupervised corpus on both the right and the left context across layers. Consequently, the pre-trained BERT model is ready to be fine-tuned by just adding a single additional layer at the output for a wide range of NLP tasks, such as question answering and language translation. Importantly, This can be achieved without substantial task-specific architecture modifications or training from the scratch. Primarily, there are two pre-trained versions of the model available; BERT-Base and BERT-Large. Being a stack of encoders, the difference between BERT-Base and BERT-Large is the number of encoder layers. In the base model, there are 12 encoder layers whereas, in the large version, it is 24.

TABLE 3. Training data used for Legal-BERT$_{base}$ [8].

Corpus	Document Count	Size (in GB)
EU legislation	61, 826	1.9 (16.5%)
UK legislation	19, 867	1.4 (12.2%)
ECJ cases	19, 867	0.6 (5.2%)
ECHR cases	12, 554	0.5 (4.3%)
US court cases	164, 141	3.2 (27.8%)
US contracts	76, 366	3.9 (34.0%)

As a result, the number of parameters or the weights and number of attention heads also differ. BERT-Large has 16 attention heads and 340 million parameters [28]. BERT-Base, a more compressed version of the same architecture has 12 attention heads with 110 million parameters. BERT-Base and BERT-Large have 768 and 1024 hidden layers corresponding to the embedding dimension respectively. Both models are pre-trained from unannotated data from the 800M words from books corpus and 2500M words from English [29]–[31]. Chalkidis *et al.* established the fact empirically that the proposed model does not generalize well in the legal domain. They came up with strategies like fine-tuning or training from scratch the BERT with domain-specific data to make the model available for specialized domains like legal which is the closest domain TI data can be compared to [8].

In the current work, we have used the pre-trained BERT model trained on 12 GB of diverse English legal corpora (Table 3) from the scratch. It has the similar network architecture as BERT-Base with 110 million trainable parameters.

The texts in the documents used in the study are longer than the maximum sequence length supported by BERT. We have produced the embedding of every chunk of text (x_i) with $32-512$ tokens long and represented the document with the mean vector (X) of all the chunks represented by the equation 3.

$$\bar{X} = \sum_{i=1}^{n} x_i \tag{3}$$

3) THE FUSION AND MLP ARCHITECTURE

We have adopted the Early Fusion approach of combining the image and text features coming from both the pre-trained

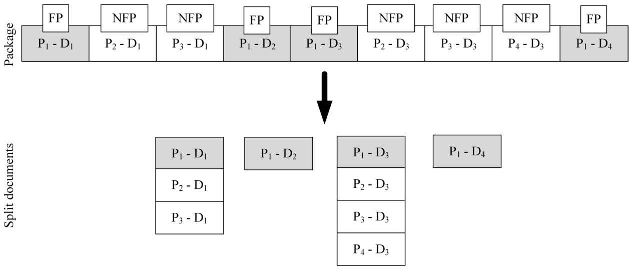


FIGURE 8. Visual representation of package segmentation algorithm.

models. The feature vector dimension of the text and output are 768 and 512 respectively. We concatenated the feature vector and passed the final feature vector into a three layered MLP with 256 units of input, 128 units of hidden and 64 units of output layer. The final output layer has a single unit with a Sigmoid activation function. Each layer has a dropout of 0.6. The overall learning rate has been fixed at $\alpha = 0.0001$. The batch size has been fixed at 50 with Adaptive Moment Optimization (ADAM) optimization function. The loss function is Binary Cross Entropy. As in the proposed approach we are depending on the transfer learning for both the image and text modalities, the training network is kept very simple. Training on too many parameters may have an catastrophic effect on the previously learnt weights on the modalities [7], [32], [33].

4) PAGE SEGMENTATION

The solution is designed after visually analyzing the documents belonging to the packages. It is observed that the first page of every document has a distinguishable textual fingerprint. The proposed data model is trained on every document's first page as a positive class and all other pages as a negative class. Once, the model which is a binary classifier, predicts and tags every page of the unseen package as a potential FP or OP of a document, Algorithm 1 is executed to segment the documents from the individual pages of the package (Figure 8 explains the splitting algorithm). The overall solution architecture is shown in Fig. 4. The same strategy is applied, and results are validated for both the package types considered for the study.

IV. EXPERIMENTS

We have apportioned the experiments into three stages. In the first stage, the experiments with CP2020 and SP2020 archive data sets using the model proposed by Wiedemann and Heyer *et al.* [13] are conducted to establish the first baseline. Experiment with the proposed architecture is carried out with the same data sets in the second stage with uni-modal networks of Image and text features independently without fusing the features as part of the ablation study. The final stage of the experiment is conducted with the proposed multi-modal network combining the image and text feature. A quantitative comparison of the result is performed to empirically establish the superiority of the proposed model over the baselines. The experiments are conducted with 5-fold cross validation for both first baseline as well as the proposed architecture. the result of the best model has been captured and reported.

Algorithm 1 DSS Algorithm Using the Output Tag of the Binary Classifier

```
Input: Package $ps_n$ where $n$ is number of pages.
Output: Set of $k$ documents, $D = \{d_k, k \leq n\}$

$ps$ = package
$i = 0$
$n$ = length of ps
$newDoc$ = empty list
$listDoc$ = empty list
while $i \leq n$ do
    if $ps[i] ==$ "FP" then
        $listDoc.append(newDoc)$
        $newDoc.clear()$
        $newDoc.add(p[i])$
    else
        $newDoc.append(p[i])$
    $i = i + 1$
listDoc.append(newDoc)
return listDoc;
```

TABLE 4. Data distribution among Training (Tr.) Validation (Val.) across all experiments. Total Document of CP2020 and SP2020 are 13, 579 and 16, 858 respectively. Total number of pages available in the data set are 79, 739 and 105, 733 for CP2020 and SP2020 respectively.

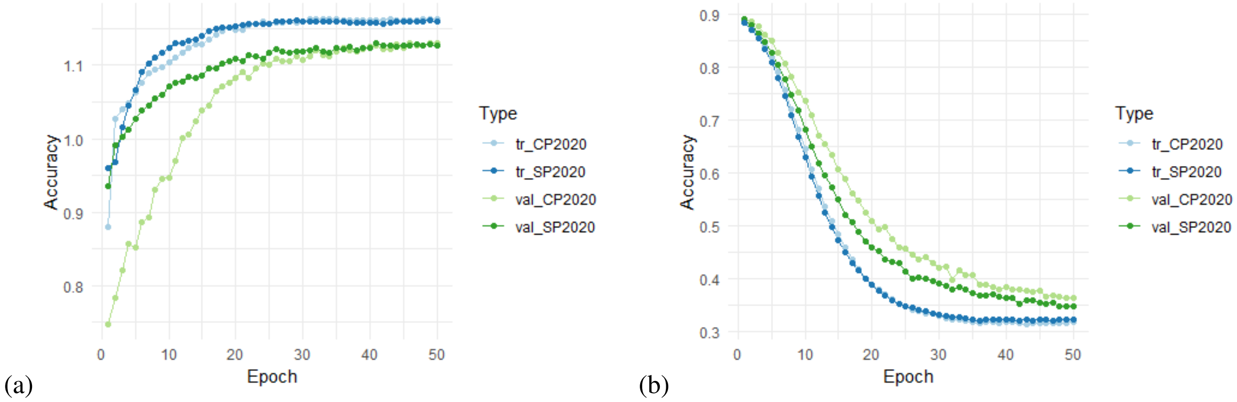
Data	CP2020		SP2020	
	Tr. (80%)	Val. (20%)	Tr. (80%)	Val. (20%)
FP	10, 863	2716	13, 486	3372
OP	52, 928	13, 232	71, 100	17, 775

In the proposed multi-modal approach, we have varied the input sequence length of the BERT encoder from 2^5 to 2^9 by incrementing the power by 1 every time keeping the embedding dimension fixed at 768 (Figure 9 and 10).

The data is split into two sets- training (Tr.) and validation (Val.) for each of the archive data set (shown in Table 4). Both CP2020 and SP2020 are randomly split into training and validation set in 80% and 20% ratio each. (Table 4).

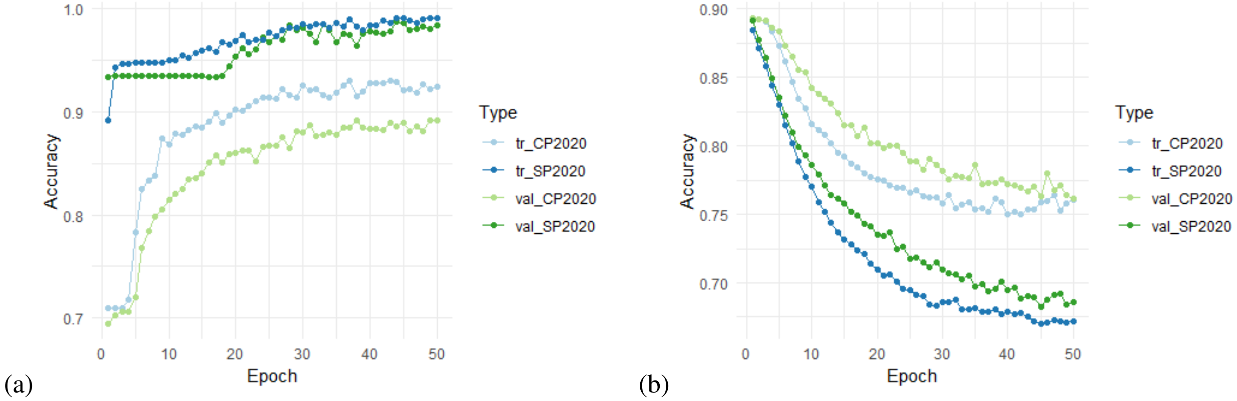
A. EVALUATION METRIC

The proposed DSS technique is based upon the paradigm of binary classification where the observations are the individual document pages. Some of the pages are FP which are considered to be the condition positive (CP) instances and OP are considered to be the condition negative (CN) observations [33]. The system predicts every observation as either FP or OP. The intersection of CP and predicted FP are termed as True Positive (TP) and remaining predicted observations are termed as False Positives (FP). Similarly, the intersection of CN and predicted OP are the True Negative (TN) cases whereas, the remaining predictions are False Negatives (FN).



(a) (b)

FIGURE 9. Baseline model training result with Legal-BERT$_{base}$ [8] for CP2020 and SP2020. (a) Training and Validation accuracies over 50 epochs on CP2020 and SP2020.(b) Training and Validation Loss over 50 epochs on CP2020 and SP2020. 'tr' represents training and 'val' represents validation.



(a) (b)

FIGURE 10. Proposed model training result with Legal-BERT$_{base}$ [8] and VGG16 for CP2020 and SP2020. (a) Training and Validation accuracies over 50 epochs on CP2020 and SP2020.(b) Training and Validation Loss over 50 epochs on CP2020 and SP2020. 'tr' represents training and 'val' represents validation.

FP and FN are also known as the Type 1 and Type 2 error respectively [34].

From the above scenario emerges the below metrics that signify the proficiency of the classifier system.

$$Accuracy = \frac{\sum TP + \sum TN}{N}, \tag{4}$$

where N = Number of total samples

$$Recall = \frac{\sum TP}{\sum CP} \tag{5}$$

Recall is also known as the True Positive Rate (TPR).

$$Precision = \frac{\sum TP}{\sum TP + \sum FP} \tag{6}$$

Due to highly imbalanced scenario of CP and CN class distribution, the Accuracy measure is also determined by a balance accuracy measure known as F_1 score.

$$F_1 = 2 \cdot \frac{\frac{\sum TP}{\sum TP+\sum FP} \cdot \frac{\sum TP}{\sum CP}}{\frac{\sum TP}{\sum TP+\sum FP} + \frac{\sum TP}{\sum CP}} \tag{7}$$

Apart from the above metrics, there is another measure which is analysed. False positive rate (FPR) is determined by

$$FPR = \frac{\sum FP}{\sum CN} \tag{8}$$

The above metrics are captured for all the proposed models and compared with the baseline model.

B. STRAIGHT THROUGH PASS (STP)

Although we depend on the accuracy of the binary classifier to gauge the efficacy of DSS, the process owners measure the effectiveness of the system by STP. This is directly related to the cost-effectiveness of the DSS. STP denotes the percentage of digital packages that need no manual intervention during the review. It is the proportion of the packages where all the pages of the stream are classified without any mistakes by the algorithm. Along with the binary classification accuracy, we have also measured and reported the STP for the experiments.

V. RESULTS AND DISCUSSION

The first baseline experiment with the model proposed by Wiedemann and Heyer *et al.* [7] fetched F_1 score of 86.76%

TABLE 5. Result of the first baseline experiment with the model proposed by Wiedemann and Heyer *et al.*

CP2020			SP2020		
F_1	ACC (%)	STP (%)	F_1	ACC (%)	STP (%)
86.76	89.31	77.15	85.33	88.54	78.32

TABLE 6. Result of the second baseline experiment with the uni-modal model only with textual feature using Legal-BERT$_{base}$.

Sequence	CP2020			SP2020		
	F_1	ACC (%)	STP (%)	F_1	ACC (%)	STP (%)
32	92.98	93.03	82.36	90.02	92.12	85.87
64	**96.37**	97.23	90.07	**97.00**	**98.48**	**93.23**
128	96.01	**97.98**	**91.21**	96.08	96.99	93.19
256	90.76	91.91	84.78	91.20	91.11	86.65
512	88.63	89.99	81.82	89.90	91.23	86.65

TABLE 7. Result of the proposed model experiment with the multi-modal features combining textual feature using Legal-BERT$_{base}$ and image feature using VGG16.

Sequence	CP2020			SP2020		
	F_1	ACC (%)	STP (%)	F_1	ACC (%)	STP (%)
32	93.43	94.39	87.38	92.89	94.23	86.31
64	**97.37**	**98.56**	**92.39**	**97.15**	**97.98**	**94.00**
128	95.39	96.12	81.23	94.83	94.33	82.12
256	87.38	88.10	76.56	86.39	88.65	79.38
512	86.56	87.03	79.87	84.09	83.67	77.08

TABLE 8. Comparative result of the proposed method with the state-of-the-art method and baseline of unimodal text feature based model using Legal-BERT$_{base}$. M_1 represents the state-of-the-art model proposed by wiedemann and Heyer *et al.* [7]. M_2 represents the unimodal model using only text modality and M_3 represents the proposed model.

Model	CP2020			SP2020		
	F_1	ACC (%)	STP (%)	F_1	ACC (%)	STP (%)
M_1	86.76	89.31	77.15	85.33	88.54	78.32
M_2	96.37	97.23	90.07	97.00	98.48	93.23
M_3	**97.37**	**98.56**	**92.39**	**97.**	**97.98**	**94.00**

with CP2020 and 85.33% with SP2020 (as reported in Table 5). The obtained STP value were 77.15% and 78.32% respectively which means 77.15% of the test documents of CP2020 and 78.32% of the test documents of SP2020 were split into individual documents with 100% accuracy and the remaining needed manual intervention. The model follows a CNN based bi-modal architecture with text and image modality. The text modality has a 350 dimension input embedding layer followed by a 1-D convolution and 256 unit dense layer. The image modality is a VGG16 pre-trained model with a cropped 256 unit final dense layer fused with each other. The result was not up to the expected accuracy set by the business process. This motivated us to experiment with the subsequent uni-modal and multi-modal models.

BERT block takes input tokens of length from 3 to 512, known as input sequence length to produce a fixed length embedding vector for the words.Texts that are longer than the input sequence length is divided into multiple such text blocks and sent to BERT. The second baseline model produced much better result than the first baseline. In this experiment, we varied the input sequence length of the text chunks from 32 to 512 and obtained the best F_1 score as well as the STP at sequence length 64 for both CP2020 and SP2020 (as reported in Table 6). The model follows an uni-modal architecture of applying only the text feature into the model. The text features are represented using the pre-trained model named Legal-BERT$_{base}$. It is a BERT family model trained on the legal domain. 12 GB of English legal text from different fields like legislation, contracts, court cases are scraped from publicly available databases. This is a light-weight version of BERT$_{base}$.

The final experiment with the multi-modal approach proposed in this work has fetched the best F_1 score and STP at sequence length 64 (as underlined in Table 7). For CP2020 data set the obtained F_1 score and STP was 97.37% and 92.39% whereas, for SP2020 it was 97.15% and 93.03% respectively.

We have achieved an F_1 score improvement of 10.61% and 11.33% over the first baseline model for CP2020 and SP2020 respectively by using our multi-modal approach. The gain in STP of the same was 15.24% and 14.71%. Comparing the proposed multi-modal approach with the uni-modal approach we see an improvement of 1.00% and 0.15% in the F_1 score; and 2.32% and 0.77% improvement in STP. Our observation is that although, the image modality alone can not perform to the desired accuracy but when combined with the text modality, it improves the overall accuracy.

Table 8 represents the comparison summary of the work where M_1 represents the model proposed by Wiedemann and Heyer *et al.* [7] and M_2 represents the model baseline unimodal model with only text features of the data using state-of-the-art transfer learning method with transformers in NLP domain. The M_3 is the proposed model with two modalities; both text and image. Marginal improvement in accuracy is clearly visible from the result that the inclusion of the image modality. However, this marginal improvement in STP can potentially save a huge cost by bypassing the human in the loop.

VI. CONCLUSION

We, in the present study, have proposed a multi-modal binary classification approach based on state-of-the-art transfer learning techniques involving images and NLP models to address the problem of DSS. Our study was motivated by a real-time need for DSS in the TI industry and a proprietary data set from a reputed TI company was considered. The previous breakthrough of the technology obtained in

2019 and 2021 have shown unsatisfactory results with our data set. The proposed multi-modal approach has been proven to have performed significantly better than the present state-of-the-art model as well as the uni-modal NLP approach combined with transfer learning (F_1 score of 97.37, 97.15 and STP of 92.39%, 94.00% for CP2020 and SP2020 archive data respectively). Output data indicate towards gain in accuracy and STP from the result obtained during the experiments. Improvement of the binary class predictability with the inclusion of the image feature has been empirically established.

Adding the image modality has only improved the model performance by 1% and 0.15% for CP2020 and SP2020 data sets. The improvements are marginal compared to the computational complexity added to the model by adding the second modality for the binary classification. However, the improvement in STP for CP2020 due to 1% F1-score upliftment is 2.32%. It cannot be ignored in a production environment where thousands of documents are processed daily by each process.

The current state-of-the-art of the Page stream segmentation task is proposed by Braz *et al.* [6] which is based on Wiedemann and Heyer et al [13]. The proposed work in our manuscript is based on Wiedemann and Heyer *et al.* [13] and it is seen that the Our proposed method on the private dataset has performed better than both the proposed model's accuracy over public datasets.

During the study we have observed the data closely and it is noticed that there are page specific visual features like the font size, margin, font type, logo presence, logo type, document title, etc. can be utilized as useful features to determine the continuation or rupture of a sequence. A deep RNN based sequence model can be trained with such features for classifying such sequences for DSS. Further research is in progress to confirm effectiveness of said sequence model.

REFERENCES

[1] A. Gordo, M. Rusinol, D. Karatzas, and A. D. Bagdanov, "Document classification and page stream segmentation for digital mailroom applications," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 621–625.

[2] A. Guha and D. Samanta, "Hybrid approach to document anomaly detection: An application to facilitate RPA in title insurance," *Int. J. Autom. Comput.*, vol. 18, no. 1, pp. 55–72, Feb. 2021.

[3] T. Meilender and A. Belaïd, "Segmentation of continuous document flow by a modified backward-forward algorithm," *Proc. SPIE*, vol. 7247, Jan. 2009, Art. no. 724705.

[4] I. Gallo, L. Noce, A. Zamberletti, and A. Calefati, "Deep neural networks for page stream segmentation and classification," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2016, pp. 1–7.

[5] A. Hamdi, J. Voerman, M. Coustaty, A. Joseph, V. P. d'Andecy, and J.-M. Ogier, "Machine learning vs deterministic rule-based system for document stream segmentation," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, Nov. 2017, pp. 77–82.

[6] F. A. Braz, N. C. da Silva, and J. A. S. Lima, "Leveraging effectiveness and efficiency in page stream deep segmentation," *Eng. Appl. Artif. Intell.*, vol. 105, Oct. 2021, Art. no. 104394.

[7] G. Wiedemann and G. Heyer, "Multi-modal page stream segmentation with convolutional neural networks," *Lang. Resour. Eval.*, vol. 55, pp. 127–150, Sep. 2019.

[8] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," 2020, *arXiv:2010.02559*.

[9] N. Audebert, C. Herold, K. Slimani, and C. Vidal, "Multimodal deep networks for text and image-based document classification," 2019, *arXiv:1907.06370*.

[10] T. Dauphinee, N. Patel, and M. Rashidi, "Modular multimodal architecture for document classification," 2019, *arXiv:1912.04376*.

[11] D. Kiela, E. Grave, A. Joulin, and T. Mikolov, "Efficient large-scale multi-modal classification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–7.

[12] R. Jain and C. Wigington, "Multimodal document image classification," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 71–77.

[13] G. Wiedemann and G. Heyer, "Page stream segmentation with convolutional neural nets combining textual and visual features," 2017, *arXiv:1710.03006*.

[14] A. Hamdi, M. Coustaty, A. Joseph, V. P. d'Andecy, A. Doucet, and J.-M. Ogier, "Feature selection for document flow segmentation," in *Proc. 13th IAPR Int. Workshop Document Anal. Syst. (DAS)*, Apr. 2018, pp. 245–250.

[15] R. Karpinski and A. Belaid, "Combination of structural and factual descriptors for document stream segmentation," in *Proc. 12th IAPR Workshop Document Anal. Syst. (DAS)*, Apr. 2016, pp. 221–226.

[16] O. Agin, C. Ulas, M. Ahat, and C. Bekar, "An approach to the segmentation of multi-page document flow using binary classification," *Proc. SPIE*, vol. 9443, Mar. 2015, Art. no. 944311.

[17] H. Daher and A. Belaïd, "Document flow segmentation for business applications," *Proc. SPIE*, vol. 9021, Mar. 2014, Art. no. 90210G.

[18] H. Daher, M.-R. Bouguelia, A. Belaid, and V. P. D'Andecy, "Multipage administrative document stream segmentation," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 966–971.

[19] M. Rusiñol, V. Frinken, D. Karatzas, A. D. Bagdanov, and J. Lladós, "Multimodal page classification in administrative document image streams," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 17, no. 4, pp. 331–341, Dec. 2014.

[20] M. A. Schmidtler and J. W. Amtrup, "Automatic document separation: A combination of probabilistic classification and finite-state sequence modeling," in *Natural Language Processing and Text Mining*. Cham, Switzerland: Springer, 2007, pp. 123–144.

[21] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," 2018, *arXiv:1805.11730*.

[22] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML*, 2011, pp. 1–8.

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

[24] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Eng. Technol. (ICET)*, Aug. 2017, pp. 1–6.

[25] P. Kim, "Convolutional neural network," in *MATLAB Deep Learning*. Cham, Switzerland: Springer, 2017, pp. 121–147.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[28] P. Ganesh, Y. Chen, X. Lou, M. A. Khan, Y. Yang, H. Sajjad, P. Nakov, D. Chen, and M. Winslett, "Compressing large-scale transformer-based models: A case study on BERT," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 1061–1080, Sep. 2021.

[29] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018, *arXiv:1801.06146*.

[30] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (Jun. 11, 2018). Improving Language Understanding With Unsupervised Learning. OpenAI. [Online]. Available: https://openai.com/blog/language-unsupervised/

[31] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 3079–3087.

[32] A. Guha, D. Samanta, A. Banerjee, and D. Agarwal, "A deep learning model for information loss prevention from multi-page digital documents," *IEEE Access*, vol. 9, pp. 80451–80465, 2021.

[33] A. Guha and D. Samanta, "Real-time application of document classification based on machine learning," in *Proc. 1st Int. Conf. Inf., Commun. Comput. Technol.*, Istanbul, Turkey: Springer, 2020, pp. 366–379.

[34] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informat.*, vol. 17, no. 1, pp. 168–192, Jan. 2021.

ABHIJIT GUHA received the B.Sc. degree (Hons.) in chemistry from Calcutta University, India, in 2006, and the Master of Computer Applications (M.C.A.) degree in computer applications from the Academy of Technology under West Bengal University of Technology, India, in 2009. He is currently pursuing the Ph.D. degree with the Department of Data Science, CHRIST (Deemed to be University), India. He is also working as a Research and Development Scientist with First American India Private Ltd. His research interests include document image processing, data mining, statistical modeling, machine learning modeling in title insurance domain, artificial intelligence, natural language processing, text mining statistical learning, and machine learning. He has delivered multiple business solutions using the AI technologies and received consecutive three ''Innovation of the Year'' awards from 2015 to 2017 by First American India for his contribution towards his research.

ABDULRAHMAN ALAHMADI received the Ph.D. degree in computer science engineering from Southern Illinois University at Carbondale, in 2019. During his studies, he was working with the Cloud Computing and Big Data Research Laboratory for a period of five years. He is currently an Assistant Professor with the Computer Science and Information Department, Taibah University, Saudi Arabia. His Ph.D. research was in cloud computing data centers scheduling for energy consumption reduction and resource utilization improvement. Since then, he has published various peer-reviewed research articles in edge and fog cloud computing. His research interests include machine learning resource management in cloud computing, task scheduling in fog computing, and the IoT-supported edge offloading techniques.

DEBABRATA SAMANTA (Member, IEEE) received the bachelor's degree (Hons.) in physics from Calcutta University, Kolkata, India, the M.C.A. degree from the Academy of Technology, under WBUT, West Bengal, and the Ph.D. degree in computer science and engineering from the National Institute of Technology, Durgapur, India, in the area of SAR image processing. He is currently working as an Assistant Professor with the Department of Computer Science, CHRIST (Deemed to be University), Bengaluru, India. He is keenly interested in interdisciplinary research development and has experience spanning fields of SAR image analysis, video surveillance, heuristic algorithm for image classification, deep learning framework for detection and classification, blockchain, statistical modeling, wireless ad hoc networks, natural language processing, and V2I communication. He has successfully completed six consultancy projects. He has received funding 5250 USD under Open Access, Publication fund. He has received funding under International Travel Support Scheme in 2019 for attending conference in Thailand. He has received Travel Grant for Speaker in Conference and Seminar, for a period of two years, from July 2019. He was invited speaker at several institutions. He is the owner of 20 patents (three design Indian patent and two Australian patent granted, and 15 Indian patents published) and two copyright. He has authored or coauthored over 177 research papers in international journal (SCI/SCIE/ESCI/Scopus) and conferences, including IEEE, Springer, and Elsevier Conference proceeding. He is the coauthor of 12 books and the co-editor of nine books, available for sale on Amazon and Flipkart. He has presented various papers at international conferences and received best paper awards. He has authored or coauthored of 21 book chapters. He is a Professional IEEE Member, an Associate Life Member of Computer Society of India (CSI), and a Life Member of the Indian Society for Technical Education (ISTE). He has received ''Scholastic Award'' at 2nd International Conference on Computer Science and IT Application, CSIT-2011, Delhi, India. He is a convener, keynote speaker, session chair, co-chair, publicity chair, publication chair, advisory board, technical program committee members in many prestigious international and national conferences. He also serves as an Acquisition Editor for Springer, Wiley, CRC, Scrivener Publishing LLC, Beverly, USA, and Elsevier.

MOHAMMAD ZUBAIR KHAN received the Master of Technology degree in computer science and engineering from U. P. Technical University, Lucknow, India, and the Ph.D. degree in computer science and information technology from the Faculty of Engineering, M. J. P. Rohilkhand University, Bareilly, India. He was the Head and an Associate Professor with the Department of Computer Science and Engineering, Invertis University, Bareilly. He has more than 15 years of teaching and research experience. He is currently an Associate Professor with the Department of Computer Science, Taibah University, Medina, Saudi Arabia. He has published more than 70 journals and conference papers. His current research interests include data mining, big data, parallel and distributed computing, theory of computations, and computer networks. He has been a member of the Computer Society of India, since 2004.

AHMED H. ALAHMADI received the Ph.D. degree in computer science and engineering from La Trobe University. His Ph.D. research was in e-health business requirements engineering. Since then, he has published various peer-reviewed research articles. He worked as the Dean of the College of Computer Science and IT, Albaha University. He is currently an Assistant Professor with the Department of Computer Science and Information, Taibah University, Saudi Arabia. He is also the Dean of the Khaybar Community College, Taibah University. In addition to research, he is also skilled in accreditation and college recruiting. He has made significant contributions in various research areas, including e-health, software engineering, business process modeling, requirements engineering, and process mining. He also has a demonstrated history of working in the higher education industry.