UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'ENERGIA DEI SISTEMI DEL TERRITORIO E DELLE COSTRUZIONI

RELAZIONE PER IL CONSEGUIMENTO DELLA LAUREA MAGISTRALE IN INGEGNERIA GESTIONALE

A Project for an Automated Identity Verification System: Development and Guidelines Definition. The Lottomatica Case

SINTESI

RELATORI

Prof. Antonella Martini
Dipartimento di Ingegneria dell'Energia dei Sistemi, del Territorio e delle Costruzioni

Ing. Alessandra Andreozzi
ELIS Consulting & Labs

Ing. Lorenzo Ricciardi Celsi
Elis Consulting & Labs

CANDIDATO

Marta Donno
marta_donno@hotmail.it

Sessione di Laurea Magistrale del 18/06/2020
Consultazione NON consentita
Anno Accademico 2019/2020

A Project for an Automated Identity Verification System: Development and Guidelines Definition. The Lottomatica Case

Marta Donno

Sommario

Al giorno d'oggi, la linea che separa il mondo fisico da quello digitale è sempre più sottile. In questo contesto, assicurare affidabilità e sicurezza nel mondo digitale, predominato dall'anonimato, è una priorità per le aziende che offrono servizi online. La presente tesi deriva dal lavoro progettuale di 5 mesi svolto all'interno del programma Junior Consulting, presso Elis Consulting & Labs a Roma. Il progetto, commissionato dal cliente Lottomatica, ha avuto l'obiettivo di creare un sistema basato sull'intelligenza artificiale (AI) in grado di eseguire automaticamente il processo di identificazione digitale degli utenti, al fine di registrarsi al Conto Gioco ed accedere alla varietà di giochi offerti da Lottomatica. Nell'ambito del progetto, il sistema è stato scomposto in tre sotto-tasks: Integrity Check, Forgery Detection e Face Detection e Matching. Per ognuno di essi questa tesi fornisce un'accurata metodologia risolutiva che fa leva sulle capacità delle reti neurali profonde di elaborare informazioni complesse sulle immagini. Inoltre, il lavoro progettuale fornisce delle linee guida sui requisiti necessari per implementare un sistema di identificazione automatico.

Abstract

Nowadays, the line separating the physical world from digital one is getting thinner. In this context, ensuring reliability and security in the digital world, dominated by anonymity, is a priority for companies that offer online services. The following thesis is the result of 5 months project carried out within the Junior Consulting program, at Elis Consulting & Labs in Rome. The work was commissioned by Lottomatica to set up an Artificial Intelligence-based system capable of performing automatically the identification process in order to register users on the online Conto Gioco and access to the variety of games offered by Lottomatica. Within the project, the system was divided into three sub-tasks: Integrity Check, Forgery Detection and Face Detection and Matching. For each of them, this thesis provides an accurate methodology that leverages the capabilities of the deep neural networks to process complex information on images. In addition, the project work provides guidelines on the requirements necessary to implement and develop an automatic identification system.

1. Context analysis

Electronic identification is referred to a digital way to demonstrate the identity of a person or organization, to execute online transactions. At present, identity documents (ID) are used as proof of identity in many online activities that include monetary transactions. However, the phenomenon of counterfeiting IDs for personal gain is increasingly frequent and fraudsters became more sophisticated. Thanks to deep learning fundamentals, the application of intelligent algorithms for image analysis is changing the way to automating internal and external processes into companies. An automatic ID verification system that uses AI affects each company level, resulting in significant benefits (Table 1).

Table 1. AI Implications on Short, Medium and Long Term

SHORT TERM	
Increase productivity	Fast document checking, simple and standard verification activities and operational efficiency increase.
Enhance reporting creation	Automatic reports creation, accurate and reliable document structured, precise process indicators are available.
Reduce fraud attempts	Wide type of tampering detection and error rates reduction. Fraudulent behavior prevention through suspicious pattern analysis.
Save costs	Operating and management costs reduction for faster verification made by the intelligent algorithm and saving in compliance cost.
MEDIUM TERM	
Compliance with local law	Ensure to meet regulatory requirement, keeping data with a real-time monitoring system.
Data integrity	Refine decisions over time for high-quality data available and automated information aggregation.
Data-driven decision	Base the decision-making process on a measured and verifiable numbers.
LONG TERM	
Increase customer experience	Immediate onboarding experience for registration time reduction and error rate abatement.
Create chat-bot services	Major customer engagement with realistic interactive interfaces that help the creation of a more human-like customer experience.

2. Problem setting

Lottomatica is an Italian company that operates globally in the gambling sector and offers the largest variety of online games to users registered in the Conto Gioco, like an electronic wallet. To be registered, users must upload an ID image, which is checked by Lottomatica to be validated. In the last five years, the increase in the number of registered users in the Conto Gioco has led to a long ID validation time. Moreover, Lottomatica complains about the growth of minors who falsify documents with editing software to access online games, resulting in a high error rate in ID validation. Lottomatica is then interested to set up an AI-based system that can perform automatically the user verification on three document types: identity card, driving license, and passport. The need arose to ensure more accurate supervision in the onboarding processes of their customers. Lottomatica has taken in charge Elis Consulting & Labs for designing and developing deep learning algorithms capable of recognizing the difference between valid and invalid ID.

3. Project scope and objectives

To obtain an automated ID verification it was necessary to split the entire process into three sub-tasks: (1) Integrity Check; (2) Forgery Detection; and (3) Face Detection and Matching. For each of them it was required to design and develop: (i) an algorithm for checking the integrity of the uploaded images; (ii) an algorithm for verifying the absence of digital manipulation; (iii) an algorithm to test that the user is the possessor of the uploaded document. The three algorithms will be then integrated into a single final system, that activates a warning on IDs images when they do not meet the pre-defined requirements.

3.1 Integrity Check, Forgery Detection, Face Detection and Matching Definition

The *Integrity Check* task is performed to certify legibility and completeness. Legibility refers to the quality of the image: it concerns the absence of blurring and movement, which are conditions that do not allow the correct reading of identification fields. Completeness refers to the layout of the document, with attention to the presence of all identification fields. *Forgery Detection* is the process aimed to determine the authenticity of the ID image, by identifying any digital modifications in the form of document size, location of the signature and other items, and content. These manipulations leave inconsistencies in the image that are invisible to the human eye, but which an intelligent algorithm can learn to recognize. Depending on the process used to produce manipulations, it is possible to identify three types of tampering[1]: (i) Splicing: a portion of an image is copied and pasted into another image; (ii) Copy and Move: a portion of an image is copied and pasted into the same image; (iii) Removal: a portion of an image is deleted. The last step of the online registration process is *Face Detection and Matching.* This sub-task aims to evaluate whether the person who claims to be the document's owner holds the ID by performing a comparison between the user selfie and the user ID. To perform this check, users must upload a second image containing their selfies with the visible document in hand.

Within the project, I worked in a team of four people. Although I worked to all of the project tasks, the activities I carried out individually were the following: collect and formalise client's requirements; develop the WBS; map the existing verification processes and the to-be solution; design a Six-Steps methodology (see Section 4) to address the Lottomatica need; prepare reports to keep track of the work through the team and externally with the client; identify the benefits associated with the project implementation; analyse the state-of-the-

[1] https://www.consilium.europa.eu/prado/it/prado-faq/prado-faq.pdf

art and support in the final architecture selection; prepare data (see sub-section 4.3.2).

4. Methodological Framework for Performing Automated Identity Verification in the Image Analytic Field

The identity verification process can be faced up by exploiting deep neural networks to address the sub-tasks of (1) Integrity Check; (2) Forgery Detection; and (3) Face Detection and Matching. To approach and solve each of these sub-tasks, I developed a methodology consisting of six-steps (see yellow boxes in Figure 1).

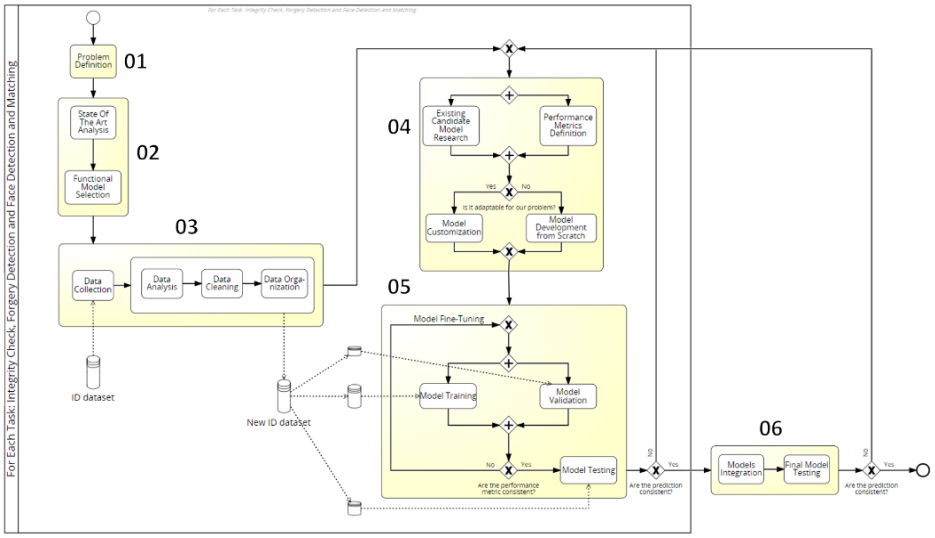

Figure 1. The Proposed Six-Steps Methodology for Designing and Developing an Automated Identity Verification System

4.1 Problem Definition

The first step is to formally define the problem to be solved. The context in which the problem to be faced lies must be well framed in an appropriate dimension. The executed tasks for automated ID verification can be treated as binary classification problems, one of the most common machine learning tasks on vector data. The model differentiates between two classes: “valid” / “invalid” or “authentic” / “manipulated” or “same person” / “different person”.

4.2 Functional Model Analysis, Evaluation and Selection

4.2.1 State of The Art

State of the art study is aimed to understand the most used and the best performing deep learning models which are already developed in the literature, together with success and failure rates. For an automatic identity verification system, the state-of-the-art analysis is reported in Section 5.2.1, Section 6.2.1, Section 7.2.1 of the thesis for Integrity Check, Forgery Detection and Face Detection and Matching tasks, respectively.

4.2.2 Functional Model Selection

The model selection activity aims to determine the final architecture to be used, given a set of candidate models. It is important to take into consideration both explicit and implicit client requirements, that open two different scenarios: (i) There exists an explicit preference for a high performing model, regardless the computational cost or complexity; (ii) There exists an explicit preference for a simple and easy-to-understand model with a low computational cost, even at the expense of its performance. Moreover, it is essential to pay attention to available resources in terms of (i) dataset; (ii) time; and (iii) technological machine. These variables influence the model's output performances.

4.3 Data Processing

4.3.1 Data Collection

The data collection activity aims to collect data to run deep learning algorithms. For the automated identification system, the dataset was not provided by Lottomatica, it is therefore necessary to download it from a public domain with open-source images. While for the Integrity Check and Face Detection and Matching tasks suitable datasets are available (MIDV 500[2] and Microsoft Celeb[3] and LFW[4], respectively), for the Forgery Detection one both complexity and lack of data led to severe limitations in completing the implementation in the time-schedule. In addition, understanding how much data is enough represents an important step in the data collection activity. In this view, one of the main issues related to the algorithm learning phase is the lack of enough data. In general, according to the amount of training data available, two macro-approaches can be distinguished[5]: (i) If a large dataset is available, simple algorithms and little hand-engineering[6] methods for data preparation and model fine-tuning are used; (ii) If a little dataset is available, a lot of hand-engineering is done to obtain good model performances. Moreover, transfer learning[7] can improve this process. In effect, to perform some of the three sub-tasks, it is necessary to do a hand-engineering and transfer learning approach for obtaining the desired performances.

[2] Arlazarov, Vladimir Viktorovich, et al. "MIDV-500: a dataset for identity document analysis and recognition on mobile devices in video stream." Компьютерная оптика 43.5, 2019.
[3] Guo, Yandong, et al. "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition." European conference on computer vision. Springer, Cham, 2016.
[4] Huang, Gary B., et al. "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments." 2008.
[5] https://www.coursera.org/learn/convolutional-neural-networks/state-of-computer-vision
[6] It is the activity of working manually on the dataset organization, on the way the data are fed into the network, and on the network parameters, due to the limited training data available.
[7] This approach became widespread through the deep neural network because the model is pre-trained on massive dataset to learn general features, allowing to obtain better performances through the transfer of knowledge from a related task with small datasets.

4.3.2 Data Preparation

The Data Preparation process consists of three activities: (1) Data Analysis; (2) Data Cleaning; and (3) Data Organization. Data Analysis is for comprehending how the recovered dataset is structured. Data Cleaning is executed for (i) removing all unwanted samples, as they are irrelevant for use; (ii) removing duplicates, as insignificant for deep neural network models. This phase can lead to a considerable loss of data. However, having representative samples of data, even if in a smaller quantity, is one of the variables that most influence model's performances. Data Organization is responsible for effectively organizing data, to use them as input for deep learning models. For each of the three sub-tasks, the final dataset is always organized into two subsets that are "valid"/"invalid", "authentic"/"manipulated" and "same person"/"different person" for Integrity Check, Forgery Detection and Face Detection and Matching, respectively.

4.4 Model implementation

4.4.1 Existing Candidate Model Research

As mentioned above, this phase is essential since many Computer Vision[8] problems have a limited amount of available data, and training a network from scratch requires adequate computational resources and time. To current knowledge, if a neural network has high performance on a problem resolution, it often works well on similar problems. For this reason, it is better to start from an existing, already developed architecture code, by using the open-source implementation as a starting point, if available. This allows faster training, validation, and testing processes.

4.4.2 Model Development

This section aims at describing the model's architectures selected for Integrity Check, Forgery Detection and Face Detection and Matching. Each one of them is composed of one or more neural network architectures, explained in the appropriate thesis Section 5.4.2, Section 6.4.2, Section 7.4.2 respectively.

4.4.3 Performance Metric Definition

In-process and post-process metrics are required in classification problems and they vary depending on the dataset and model's architecture. In fact, each of the three sub-tasks uses

[8] Computer Vision is an interdisciplinary scientific field that seeks to develop models from which computers can gain high-level understanding from digital images.The ultimate goal of computer vision is to mimic human visual perception.

different evaluation metrics[9]: (i) Precision, Recall and Area Under the Curve (AUC) for the Integrity Check task; (ii) Mean Average Precision (mAP), Receiver Operating Characteristic (ROC) Curve, Area Under the Curve (AUC) for the Forgery Detection task; (iii) Accuracy for the Face Detection and Matching task.

4.5 Model Training, Validation and Testing

For performing training, validation and testing it is necessary to split the dataset into three parts (see Figure 2 in Appendix A). The *training data* represent the sample of the dataset used to train the model. The model learns from this data. The *validation data* are used to fine-tune the model hyperparameters, allowing to check the progress of the model in-process. These two phases take place simultaneously. Before starting a training procedure with deep neural networks, it is necessary to determine parameters: (i) Number of epochs[10]; (ii) Batch size[11]; (iii) Early stopping patience[12]; (iv) Number of classes[13]. During these phases, the training and validation loss function trends are monitored to avoid overfitting[14], which occurs when the function trend is divergent, and to control they are decreasing, which means the network is learning. Finally, the *testing data* are used to provide an unbiased evaluation of a final model. These data have never been seen by the model therefore, they are completely new. The test set generates the final model evaluation metrics.

4.6 Models Integration and Final Testing

Finally, the three sub-tasks must be integrated for generating a final output that ensures user identification. In this case, integration occurs simply by recalling the models in sequence. To use this integrated system, the company requires to load it on a Docker platform[15] integrated into the online site. To start the automatic verification the user must upload: (i) The photo of the ID; (ii) The selfie with the ID in the hand.

5. Results

Six methodological steps are applied specifically for the sub-tasks of Integrity Check, Forgery Detection and Face Detection and Matching, as listed in Table 2, Table 3 and Table 4 respectively. High performances for the Integrity Check task were achieved, reaching the

[9] Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." 2011.
[10] How many times the algorithm sees the training set of data considering forward and backward propagation.
[11] The number of training examples in a forward/backward step.
[12] The number of epochs to wait before early stop the validation set if there is no progress in loss performance.
[13] The number of homogeneous groups in which the deep learning models can classify the input images.
[14] The model does not generalize well from the training data to unseen data.
[15] A Docker is a complete platform which allows to contain the deep learning model.

following results: Precision=94.96%, Recall=94.78%, AUC=98.16%. Finally, for Face Detection and Matching, Accuracy of 87.5% was obtained, allowing adequate facial matching.

Table 2. Phases, Sub-phases, Key Facts, Results and Thesis Paragraphs for Integrity Check Task

Phase	*Sub-phase*	*Key Facts*	*Results*	*§*
Problem Definition		- Binary classification problem in a Supervised Learning; - Check document legibility and completeness.	Image classification: • 1: valid image; • 0: invalid image.	5.1
Functional Model Analysis, Evaluation and Selection	State of the art	- Integrity Check is not addressed in image analytics field; - Visual-base method for document classification can be adapted to the Integrity Check problem.	Three-steps approach: - Local features extraction; - Local features aggregation into a global descriptor; - Global image descriptor classification.	5.2
	Functional model selection	- Selected models allow checking the legibility and the completeness of the uploaded image.	- Quality estimator; - Segmentation; - NetVLAD.	
Data Processing	Data collection	- Open-source dataset available.	Dataset MIDV-500.	5.3
	Data preparation	- Dataset analysis; - Dataset cleaning by removing documents with different alphabets, layout or different type of document; - Dataset organization into two folders: • Valid: images conform to 6 requirements (no cut, no reflections, no objects above, no blurring, every number/letter is clear, no interpretations); • Invalid: the other ones.	A clean and well-organized dataset into two labelled folders: • Valid; • Invalid.	
Model Implementation	Existing candidate models research	- Hybrid approach is used: • transfer learning; • customizing existing models; • implementing other model from scratch.	- Quality estimator: trained via transfer learning on ImageNet dataset; - Segmentation: customized on code implementation available; - NetVLAD architectures: implemented from scratch.	5.4
	Model development	- Cascaded approach; - Write in Python 3.6; - Libraries used for code implementation: Numpy, TensorFlow 2.0, Tensorboard.	- Quality estimator: Mobile Net; - Segmentation: U-Net; - NetVLAD: cropped VGG-16, NetVLAD pooling layer, PCA and a Fully Connected Network for classification.	
	Performance metric definition	- High performance required to avoid advancing an unreadable or incomplete photo that has strong repercussion on recognizing tampering; - Monitor True Positives[16] (TP).	Metrics: • Recall • Precision • AUC	
Model Training, Validation and Testing		- Each architecture is trained, validated and tested individually.	- 2139 training images, 611 validation image, 200 epochs, 2 classes and 32 as batch size; - 306 test images: • Precision: 94.96% • Recall: 94.78% • AUC: 98.16%	5.5
Models Integration and Final Testing		- Integration in sequence through a model recall; - Testing of the final output.	Integrity Check final single model.	5.6

[16] TP: the image is classified as "valid" and in reality it is "valid".

Table 3. Phases, Sub-phases, Key Facts, Results and Thesis Paragraphs for Forgery Detection Task

Phase	*Sub-phase*	*Key Facts*	*Results*	*§*
Problem Definition		- Object detection and localization problem.	Object detection and localization: - Detect forgeries in the image; - Localize in which pixels the manipulation occurs.	6.1
Functional Model Analysis, Evaluation and Selection	State of the art	- Forgery Detection is not a solved problem in the scientific community, it is in an experimental phase; - Graph-base method for forgery localization.	Three-steps approach: - Split the image into patches; - Compute a similarity score between patches; - Create a cluster-base graph.	6.2
	Functional model selection	- The models selected allow to check if an image is manipulated and localize the manipulation.	- Forensics similarity score; - Graph with partitioned communities according to the similarity score.	
Data Processing	Data collection	- Dataset requirements: • Images with all the manipulation cases (Splicing, Copy and Move, Removal); • Images with labelled manipulation.	Datasets MIDV-500 and DRESDEN[17].	6.3
	Data preparation	- DRESDEN dataset analysis; - Synthetic Dataset creation; - Dataset organization into two folders: • Manipulated; • Authentic.	A synthetic and well-organized dataset into two labelled folders: • Manipulated; • Authentic.	
Model Implementation	Existing candidate models research	- Model implementation from scratch.	- Forensic Similarity; - Forensic Similarity Graph.	6.4
	Model development	- Cascaded approach.	- Forensic Similarity: MISLnet and a 3-layer similarity network; - Forensic Similarity Graph: graph construction with vertex (patches) and weighted edges (similarity score).	
	Performance metric definition	- Reduce False Positive[18] (FP) and False Negative[19] (FN).	Metrics: • mAP • ROC • AUC	
Model Training, Validation and Testing		Guideline: - Train, validate and test phase are performed in sequence (as the output of the Forensic Similarity is the input of the Forensic Similarity Graph).	Not performed yet for three factors: • Time constraints; • Lack of data available; • Technological constraints.	6.5
Models Integration and Final Testing		Guideline: - Integration in sequence through a model recall; - Testing of the final output.	Forgery Detection final single model.	6.6

Table 4. Phases, Sub-phases, Key Facts, Results and Thesis Paragraphs for Face Detection and Matching Task

Phase	*Sub-phase*	*Key Facts*	*Results*	*§*
Problem Definition		- Face verification problem.	Face verification: • 1: faces belong to the same person; • 0: faces belong to different person.	7.1
Functional Model Analysis, Evaluation and Selection	State of the art	- Face Detection and Matching is a fully solved problem in face recognition field.	Three-steps approach: - Face and landmark detector; - Face alignment; - Deep features extraction and face matching.	7.2
	Functional model selection	- The models selected allow to check if the person who claims to be the owner	- Face and landmark detection;	

[17] Gloe, Thomas, and Rainer Böhme. "The dresden image database for benchmarking digital image forensics." Journal of Digital Forensic Practice 3.2-4: 150-159, 2010.
[18] FP: the image is classified as "manipulated", but it is "authentic".
[19] FN: the image is classified as "authentic", but it is actually "manipulated".

		of the document is its possessor.	- Face embedding; - Classification.	
Data Processing	Data collection	- Dataset requirements: • Images with faces in different angles; • Images with faces in different facial expressions.	Datasets Microsoft Celeb and LFW.	7.3
	Data preparation	- Microsoft Celeb and LFW dataset analysis; - Microsoft Celeb and LFW dataset cleaning; - Selfie-ID dataset creation; - Dataset organization: • A folder for each person in the dataset; • Images associated with the person in each folder.	A cleaned and well-organized dataset into *n* labeled folders, containing different images with different people expressions and angles.	
Model Implementation	Existing candidate models research	- "Ready-to-use" and available model implementation.	- Face and landmark detection; - Face embedding; - Classification.	7.4
	Model development	- Cascaded approach; - Write in Python 3.6; - Libraries used for code implementation: Numpy, OpenCV 2.0, PyTorch.	- Face and landmark detection: Multi-Task Cascaded Neural Network; - Face embedding: face alignment, image flip and merge, ResNet CNN; - Classification: cosine similarity.	
	Performance metric definition	- High accuracy in matching performances.	Metric: • Accuracy	
Model Training, Validation and Testing		- Each architecture is already trained and validated in the model's implementation available; - The test is performed individually.	- Over 4M training and validation images from Microsoft Celeb and LFW dataset; - 100 test images: • Accuracy: 87.5%	7.5
Models Integration and Final Testing		- Integration in sequence through a model recall; - Testing of the final output.	Face Detection and Matching final single model.	7.6

6. Benefits

6.1 As-is Process: Manual Identity Verification

In the as-is ID verification process (see Figure 3 in Appendix B), Lottomatica sends the image uploaded by the user to Comdata, an external society aimed at verifying the customer's ID image. If the image presents some legibility problems, Comdata informs Lottomatica to contact the user for uploaded another image. In the case of suspected manipulation, Comdata will send the image to the Anti-Fraud Unit, which is a Lottomatica division skilled for detecting forgeries. The entire checking process is disposed to mistake as (i) Limitations in scale, both in terms of processing time and costs; (ii) Sophisticated manipulations are not visible to the human eye, therefore several False Negatives is generated; (iii) Management and operative cost associated with internal and external relation is more than a machine, especially at high-scale; (iv) Long waiting times for user registration.

6.2 To-be Process: Automated Identity Verification

The application of an automatic system to verify ID images (see Figure 4 in Appendix B) has a significant impact on Lottomatica processes. In this view, it can benefit from (i) Saving in the use of resources, (i.e., minimization of intervention and elimination of Comdata checks); (ii) Reduce management and operating costs associated with the external service provider; (iii) Manage huge sets of images, accelerating the verification process; (iv) Detect certain manipulations that human capabilities cannot; (v) Increase the accuracy of the document verification procedure; (vi) Increase user experience by saving registration time.

7. Conclusions and Future Develoments

This thesis presents a customized framework for approaching the identification problem through data science techniques. In addition, the goal was to provide guidelines for managers and practitioners involved in the design and development of deep learning-based prototypes. Designing an automated system for identity verification by a formalized framework allows to systematically approach a complex case, in a constantly evolving context from a scientific and technological point of view. Moreover, an automated user verification in the gambling value chain's processes is an effective way to ensure safety and reliability, as well as increasing customer experience. The selected architectures for the Integrity Check, Forgery Detection, and Face Detection and Matching sub-tasks are a starting point to test the best configurations that can achieve high performance in replacing manual verification activity. In this view, the usage of deep learning-based systems is the right direction to implement a fully digitized system. To compare these sub-tasks, a multi-criteria approach has been used, as shown in Appendix C, Table 5. In the presented case, the six-steps methodology opened scenarios that can be explored to further increase the benefits deriving from processes automation. Three main possible future scenarios are the following:

- Implementing and testing the proposed architectures for Forgery Detection task;
- Decoupling image acquisition from ID verification to reduce management and storage costs through a solution which completes the ID verification on the customer device;
- Implementing a user's real-time video system to prevent the fake-faces problem.

***Dissemination*:**

Donno, M., Andreozzi, A., Martini, A. (2020), "*Design of a Methodology for Automated Identity Verification: The Case of the Gambling Industry*"**,** submitted to itAIS - XVII Conference of the Italian Chapter of AIS Organizing in a digitized world: Diversity, Equality and Inclusion, Chieti-Pescara, 16-17 October

APPENDIX

Appendix A. Model Training Validation and Testing: dataset split ratio

To perform training, validation, and testing operations the dataset was split into three parts depending on the amount of data available, as shown in Figure 2 below.

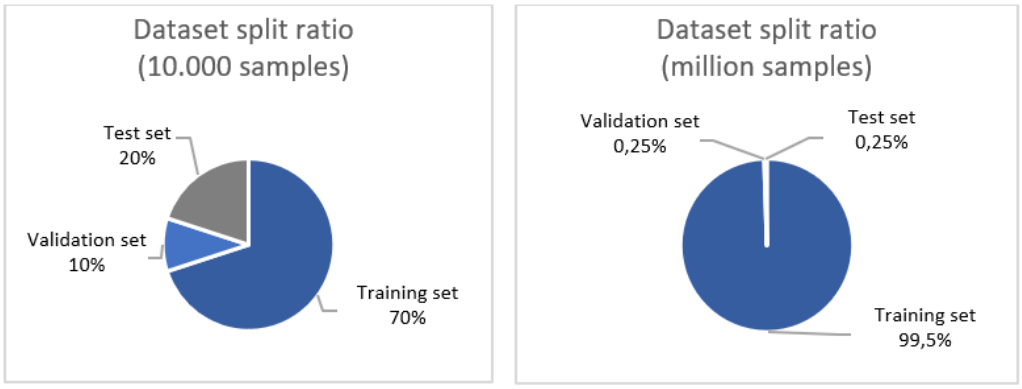

Figure 2. Dataset Split Ratio. From left to right: Ratio for Dataset within 10.000 Samples and Ratio for Dataset with Million Samples

Appendix B. From Manual to Automated Identity Verification Process: as-is and to-be process

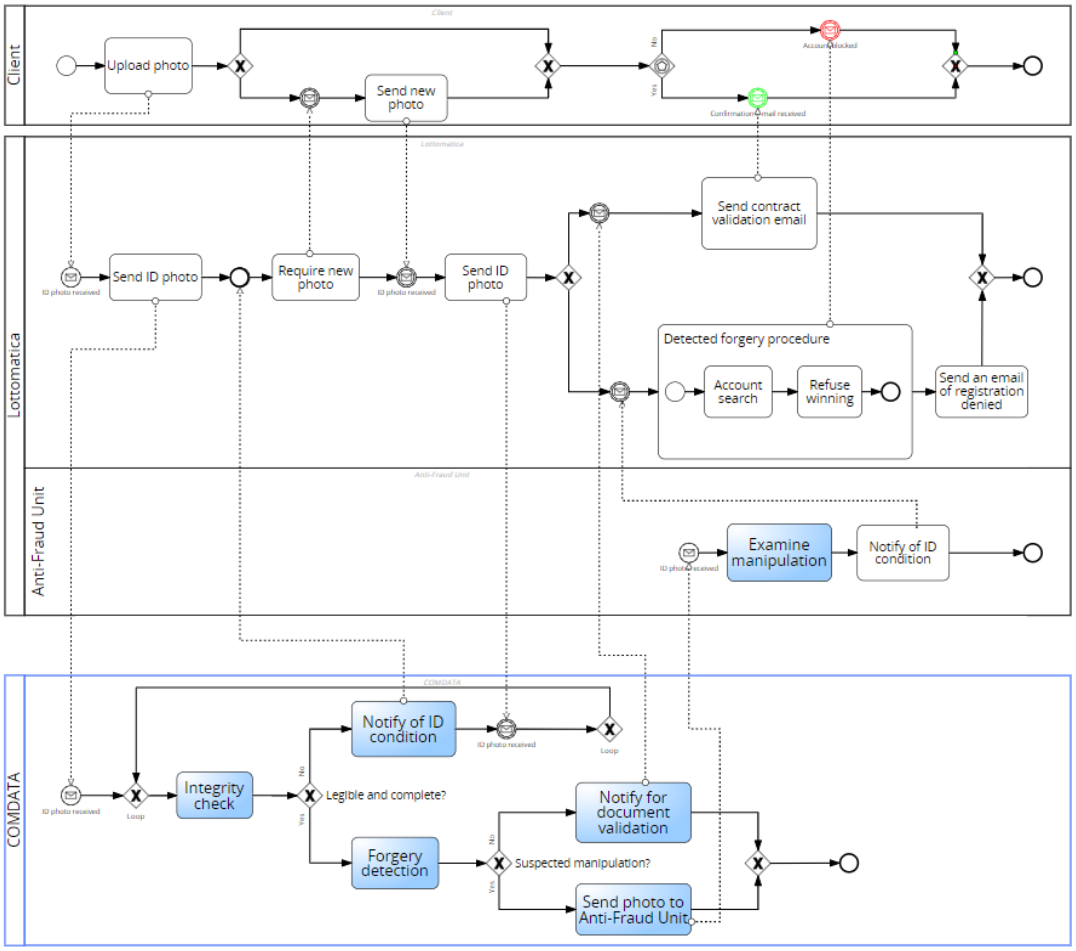

Figure 3. Identity Verification Process: as-is Manual Verification

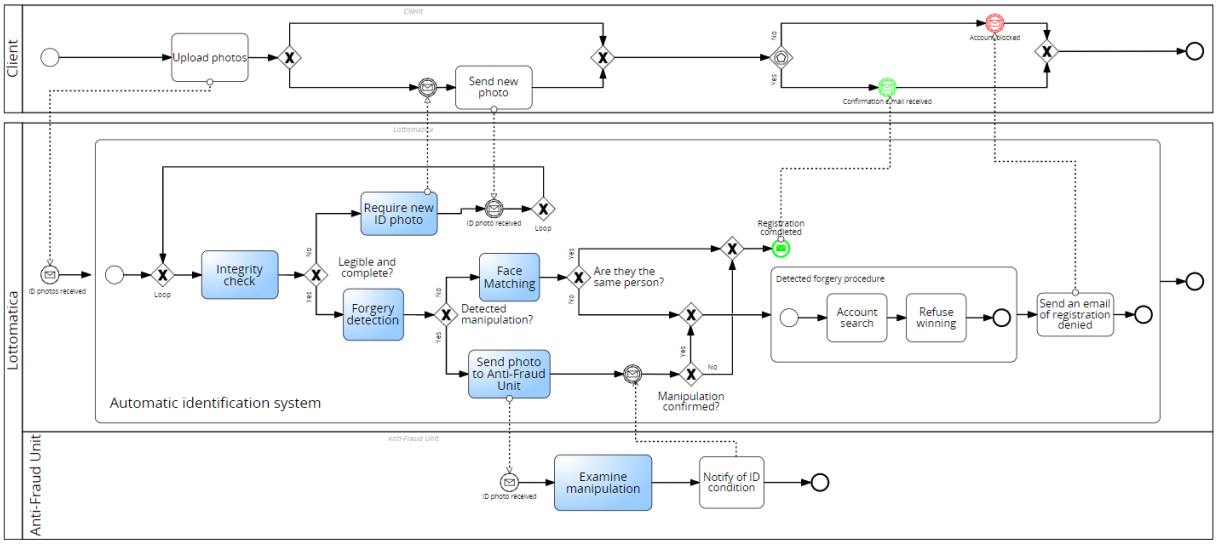

Figure 4. Identity Verification Process: to-be Automated Verification

Appendix C. Conclusions: Lesson Learned and Future Developments

Table 5. A Multi-Criteria Comparison for Integrity Check, Forgery Detection and Face Detection and Matching

CRITERIA / TASK	CRITERION 1 *Problem addressed by the scientific community*	CRITERION 2 *Problem solved by the scientific community and well documented in literature*	CRITERION 3 *Data-set availability and its access*	CRITERION 4 *Code implementation availability*
INTEGRITY CHECK	No	No	Yes	Partially
FORGERY DETECTION	Yes	No	No	No
FACE DETECTION AND MATCHING	Yes	Yes	Partially	Yes

Table 5 shows that the Integrity Check task has not been taken into consideration in the Image Analitycs field. Fortunately, an appropriate open-source dataset is available. The accessibility to this resource has strongly contributed to the success of the Integrity Check model, implementing a solution never explored so far and highly performing. Forgery Detection, on the other hand, is a task addressed but still in the experimental phase and therefore not completely solved in the scientific community. The biggest challenge in approaching these tasks is the lack of a suitable dataset to train deep learning models. On the other hand, Face Detection and Matching is considered a widely solved task within the face verification field and this was the key success driver for delivering the results reported in Table 4.

Appendix D. La mia esperienza in ELIS Consulting & Labs

Questo lavoro di tesi è il frutto di un progetto svolto nell'area Technology & Innovation di Elis, durante il percorso di 5 mesi del programma Junior Consulting. Sono molto interessata al mondo della consulenza e alla figura del Data Scientist, che richiede un mix di abilità tecniche e di soft skills, ed è proprio per questo che ho scelto Junior Consulting. Sono venuta a conoscenza del programma formativo JC di Elis grazie alla Prof.ssa Antonella Martini dell'Università di Pisa. Già dal primo giorno, ho compreso l'unicità di questo percorso che, attraverso una formazione mirata, mi ha dato l'opportunità di lavorare su un progetto stimolante in ambito di Intelligenza Artificiale (AI), aiutandomi ad accrescere, al tempo stesso, le mie competenze trasversali attraverso attività di Team Building, Project Management e di Comunicazione con i Top Manager delle aziende che fanno parte del network di Elis. Il percorso JC prepara le persone affinché possano fare la differenza, in un ambiente giovane, dinamico e concreto in cui si pone l'accento sulla valorizzazione e sulla crescita personale e professionale. Durante questi mesi infatti, ho compreso pienamente cosa significa la parola *collaborazione,* attraverso la partecipazione attiva al progetto per Lottomatica in un team eterogeneo dal punto di vista del background tecnico. Il team con il quale ho lavorato era composto da Lorenzo (Team Leader, Ing. Robotico e Dell'Automazione), Luca (studente di dottorato, Ing. Informatico) e Valerio (Ing. Robotico e dell'Automazione). La collaborazione si è esplicitata nel lavoro giornaliero di formazione continua, condividendo le informazioni e le risorse sullo stato di avanzamento lavori. Sono molto felice di aver acquisito nozioni e competenze tecniche riguardanti la progettazione e lo sviluppo di modelli di AI, comprese le competenze di programmazione che porterò con me nel mio futuro professionale. Il mio ruolo ha incluso anche l'attività di traduzione delle scelte tecnico-implementative in report aziendali che potessero essere utili sia internamente, per tener traccia del lavoro svolto, sia esternamente per la comunicazione con il cliente. Per concludere, ringrazio Elis ed i miei compagni di team per avermi dato l'opportunità di lavorare con persone uniche con le quali ho condiviso momenti di difficoltà e di stallo, ma soprattutto momenti di crescita, di divertimento e di euforia per i risultati raggiunti.

Figura 5. Io, Luca e Valerio al Kick-off di progetto nella sede di Lottomatica

Tabella 6. Competenze tecniche e strumenti acquisiti nel percorso di tesi in Elis Consulting & Labs

Fase del progetto	***Competenze tecniche acquisite***	***Strumenti***
State of the art Analysis	Overview sui modelli algoritmici allo stato dell'arte per svolgere operazioni di *Image Classification, Forgery Detection, Object Detection and Localization and Face Verification.*	Google Scholar, Scopus
Data Preparation	Raccolta, analisi, pulizia e organizzazione dei dataset necessari per la fase di training dei modelli di rete.	Python, libreria OpenCV
Implementation Methodology	Esecuzione del flusso di fasi per l'implementazione di un'architettura di rete, transfer learning, implementazione from scratch.	Jupyter Notebook, Google Colab
Algorithms Training, Validation and Testing	Implementazione delle funzionalità di classificazione di una immagine tramite reti neurali profonde.	Python, librerie Tensorflow 2.0, Pytorch, Numpy