[19] Patents Registry
The Hong Kong Special Administrative Region
香港特別行政區
專利註冊處

[11] 30057394 A

[12]

SHORT-TERM PATENT SPECIFICATION
短期專利說明書

[21] Application no. 申請編號
32021043681.4

[51] Int. Cl.
G06F 19/00 (2011.01)

[22] Date of filing 提交日期
03.12.2021

[45] Date of publication of grant of patent 批予專利的發表日期
08.04.2022

[73] Proprietor 專利所有人
City University of Hong Kong
Tat Chee Avenue, Kowloon
HONG KONG

[72] Inventor 發明人
Kei Hang Katie CHAN 陳紀行
Yat Ming WOO 胡日明
Jundong LIU 劉俊東
Ruixuan HUANG 黃睿軒
Tsz Kin WAN 温梓健

[74] Agent and / or address for service 代理人及/或送達地址
IDEA INTELLECTUAL LIMITED
21st Floor, Skyway Centre
23 Queen’s Road West, Sheung Wan
HONG KONG

[54] METHOD AND SYSTEM FOR MACHINE LEARNING AND DEEP LEARNING BASED ASSESSMENT OF STROKE
基於機器學習和深度學習的中風評估方法和系統

[57] A machine learning and deep learning-based stroke assessment system, comprising: a clinical data collection module configured to extract and organize stroke patients related clinical data into structured, textual, and image datasets; a machine learning and deep learning-based mortality prediction module configured to analyse the structured and the textual datasets to predict a long- or short-term mortality rate of the stroke patients; a machine learning-based accurate stroke subtype classification module configured to analyse the structured datasets to generate a classification of stroke subtypes suffered by the stroke patients; a machine learning and deep learning-based treatment decision module configured to analyse the structured, textual, and image datasets to generate medical decisions for treating the stroke patients; and a deep learning-based large vessel occlusion (LVO) patient classification module configured to analyse the structured and image datasets to screen out LVO patients from ischemic stroke patients.

一種基於機器學習和深度學習的中風評估系統，包括：臨床數據採集模塊，被配置為中風患者相關的臨床數據提取並組織成結構化、文本化和圖像化的數據集；基於機器學習和深度學習的死亡率預測模塊，被配置為分析結構化和文本數據集，以預測中風患者的長期或短期死亡率；基於機器學習的準確中風亞型分類模塊被配置為分析結構化數據集以生成中風患者所患中風亞型的分類；基於機器學習和深度學習的治療決策模塊，被配置為分析結構化、文本和圖像數據集，以生成治療中風患者的醫療決策；以及基於深度學習的大血管閉塞 (LVO) 患者分類模塊，被配置為分析結構化和圖像數據集，以從缺血性中風患者中篩選出 LVO 患者。

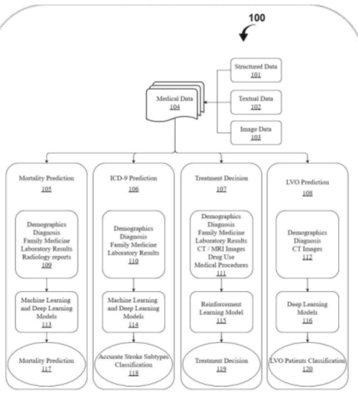
100
Structured Data 101
Textual Data 102
Image Data 103
Medical Data 104
Mortality Prediction 105
ICD-9 Prediction 106
Treatment Decision 107
LVO Prediction 108
Demographics Diagnosis Family Medicine Laboratory Results Radiology reports 109
Demographics Diagnosis Family Medicine Laboratory Results 110
Demographics Diagnosis Family Medicine Laboratory Results CT / MRI Images Drug Use Medical Procedures 111
Demographics Diagnosis CT Images 112
Machine Learning and Deep Learning Models 113
Machine Learning and Deep Learning Models 114
Reinforcement Learning Model 115
Deep Learning Models 116
Mortality Prediction 117
Accurate Stroke Subtypes Classification 118
Treatment Decision 119
LVO Patients Classification 120

FIG. 1

METHOD AND SYSTEM FOR MACHINE LEARNING AND DEEP LEARNING BASED ASSESSMENT OF STROKE
基於機器學習和深度學習的中風評估方法和系統

Inventors: Kei Hang Katie CHAN, Yat Ming WOO, Ruixuan HUANG, Jundong LIU, and Tsz Kin WAN

Field of the Invention:

[0001] The present invention relates to the application of machine learning and deep learning-based methods and systems for assessing and predicting strokes in a population using medical data.

Background of the Invention:

[0002] Strokes are some of the leading causes of death and disability across the globe. There are 13.7 million new stroke cases worldwide every year, and 5.5 million people die from strokes each year. Moreover, strokes are some of the primary causes of disability and death worldwide. The Hong Kong Special Administrative Region (HKSAR) is facing a significant challenge in relation to strokes as a group it is currently the 4th leading cause of death in HKSAR. Each year, there are approximately 25,000 new stroke patients, accounting for 0.8% of its population. Specifically, every year, 3.2% of individuals aged 65 years and older suffered from strokes, which resulted in more than 1,000 deaths as direct causes.

[0003] A stroke occurs when a blood clot blocks the blood vessels that carries oxygen and nutrients to the brain. As a result, brain cells are irreversibly damaged and begin to die within a few minutes. Thus, stroke patients' mortality and disability rates are very high. There are two sub-types of strokes: ischemic strokes and haemorrhagic strokes. Ischemic strokes account for more than 60% of stroke cases. They are due to blood vessels in the brain becoming too narrow or blocked by plaques and thereby result in insufficient blood supply to the brain. In particular, large vessel occlusions (LVOs) are a subtype of ischemic stroke where either the proximal intracranial anterior or the posterior circulation are obstructed. LVO can cause a variety of serious neurological symptoms, including facial paralysis, dyskinesia, and language disorders. Haemorrhagic strokes, however, account for approximately 30% of stroke cases and are due to the rupture of blood vessels resulting in blood flooding into the brain. This causes severe damage to brain cells and can even result in cell death.

[0004] Medical data is diverse and highly complex. The use of machine learning and deep learning methods to analyse medical data including structured, text, and images could better predict and help prevent strokes more effectively and thereby provide stroke patients with a more accurate diagnosis and more effective treatment. Therefore, a better stroke assessment system based on machine learning and deep learning is wanted in the art. Such stroke assessment system would be beneficial to both medical providers and patients as it can provide doctors with effective diagnostic support, medical decision-making assistance, and in turn provide better services for stroke patients as a result.

Summary of the Invention:

[0005] It is an objective of the present invention to provide a stroke assessment system based on machine learning and deep learning that addresses the aforementioned needs in the art. In accordance with one aspect, the stroke assessment system comprises a clinical data collection module, a mortality prediction module, an accurate stroke subtypes classification (ICD-9 prediction) module, a treatment decision module, and a LVO patient’s classification module.

[0006] In accordance with various embodiments, the stroke assessment system uses structured data, image data, textual report data as input to each of the modules, through data pre-processing and data analysis steps, to complete the function of each of the module.

[0007] The mortality prediction module predicts the mortality of one or more stroke patients. The module receives structured data (including diagnosis information, biological test data, etc.) and textual data (including patient clinical reports, radiology test reports, etc.) of the stroke patients as the module's input, processes the structured data through one or more networks of multi-level machine learning model, deep learning model, and/or ensemble models thereof, then generates and outputs short-term or long-term probabilities of deaths of the stroke patients.

[0008] The accurate stroke subtypes classification module predicts specific subtypes of stroke for the stroke patients. The module receives clinical data of the stroke patients as the module's input, processes the clinical data through one or more networks of machine learning model and deep learning model, then generates and outputs specific subtypes of the stroke.

[0009] The treatment decision module generates medication and treatment recommendations for the stroke patients. The module receives basic physical information, biological test data, radiology test data, and textual reports of the stroke patients as input, processes these data through a deep learning network, then generates and outputs the medication and treatment recommendations that enable the patient to obtain the best prognostic effect.

[0010] The LVO patient classification module is an end-to-end LVO screening tool based on a deep learning model. The module receives radiological examination data and clinical data of one or more of the stroke patients identified by ischemic strokes as input, processes the radiological examination data and clinical data through a deep learning network, generates and outputs a probability of each of the stroke patients being a LVO patient.

Brief Description of the Drawings:

[0011] The scope of the present invention is not limited to the content of the accompanying drawings provided. In order to make the advantages of the present invention easier to understand, aspects and embodiments of the present invention are illustrated in the drawings, in which:

[0012] FIG. 1 illustrates the overall structure of a stroke assessment system in accordance with one embodiment of the present invention;

[0013] FIG. 2 illustrates the data sources required by the stroke assessment system;

[0014] FIG. 3 illustrates the pre-processing pipeline of structured data used in the stroke assessment system;

[0015] FIG. 4 illustrates the pre-processing pipeline of textual data used in the stroke assessment system;

[0016] FIG. 5A and FIG. 5B illustrate the pre-processing pipeline of image data used in the stroke assessment system;

[0017] FIG. 6 illustrates the machine learning and deep learning models' training, validating, and testing pipeline for structured data in accordance with one embodiment of the present invention;

[0018] FIG. 7 illustrates the machine learning, deep learning and ensemble models' training, validating, and testing pipeline for structured and textual data in accordance with one embodiment of the present invention;

[0019] FIG. 8 illustrates the deep learning model's training, validating, and testing pipeline for structured and image data in accordance with one embodiment of the present invention; and

[0020] FIG. 9 illustrates the machine learning, deep learning, and reinforcement learning models' training, validating, and testing pipeline for structured data, image data and textual data in accordance with one embodiment of the present invention.

Detailed Description:

[0021] In the following detailed description, the figures are referred to in explaining the detailed designs of the various embodiments of the present invention. Certain figures and descriptions provide examples of some modules and models in the stroke assessment systems in accordance with the various embodiments, so that ordinarily-skilled persons in the art can better understand the present invention, better adopt its embodiments, and use them more effectively. It should be understood that various implementations and logical and structural modifications to the modules of the stroke assessment system are readily realizable without undue experimentation and departure from the spirit of the present invention.

[0022] The following detailed description is intended to provide an exemplary description of the present invention, rather than a restrictive description. When introducing the stroke assessment system and its modules, the articles "a", "an", and "the" are intended to mean that there are one or more elements. The terms "comprising", "including", and "for example" are intended to be inclusive and mean that there may be other elements in addition to the listed elements.

[0023] In the referenced figures, there are different modules and their function realization processes. The various elements in the figures are connected by arrows or surrounded by boxes. This does not mean that the related modules have a geographical relationship, but instead the arrows indicate the logical relationship between the boxed content. These connections can be realized by a corresponding application, a piece of code, or commands executed by multiple pieces of code stored in multiple memories.

[0024] All or portions of the embodiments disclosed herein may be implemented using one or more of specially configured computing devices, computer processors, or electronic circuitries including but not limited to graphics processing units (GPUs), application specific integrated circuits (ASICs), field programmable gate arrays (FPGAs), and other programmable logic devices configured or programmed according to the teachings of the present disclosure. Computer instructions or codes running in the computing devices, computer processors, or programmable logic devices can readily be prepared by practitioners skilled in the software or electronic art based on the teachings of the present disclosure. The aforesaid one or more computing devices may include one or more of server computers, personal computers, laptop computers, mobile computing devices such as smartphones and tablet computers.

[0025] The electronic embodiments include computer-readable storage media having the computer instructions or codes stored therein, which can be used to configure or program the computing devices, computer processors, or electronic circuitries to perform any of the processes of the present invention; and to store data generated by any of the processes of the present invention. The computer-readable storage media include, but are not limited to, floppy disks, optical discs, Blu-ray Disc, DVD, CD-ROMs, magneto-optical disks, solid-state discs, ROMs, RAMs, SRAMs, DRAMs, flash memory devices, electrically programmable read-only memories (EPROMs), electrically erasable programmable read-only memories (EEPROMs), or any type of media or devices suitable for storing instructions, codes, and/or data.

[0026] Various embodiments of the present invention also may be implemented in distributed computing environments and/or Cloud computing environments, wherein the whole or portions of computer instructions or codes are executed in distributed fashion by one or more processing devices interconnected by a communication network, such as an intranet, Wide Area Network (WAN), Local Area Network (LAN), the Internet, and other forms of data transmission medium.

[0027] The programming language used to implement the embodiments disclosed herein is not limited, and the functions of each module in the systems in accordance with the various embodiments can be implemented by one or more programming languages, such as R, C, C++, Python, Java, etc.

[0028] The data processing described herein is only used as a sample, and its description is not used as a data use restriction of the present invention. Data types and data contents other than the data sample also have the possibility of being loaded by the various embodiments of the present invention and run smoothly.

[0029] The numbers in the boxes of each element in the system are only used to make the explanation of the systems in accordance with the various embodiments more convenient. In all the figures, the same number represents the same element, including alternative embodiments of the same element. Numbers are only used for labelling, and the size of their value does not represent the relationship between elements.

[0030] Referring to FIG. 1 for the following description. In accordance with one embodiment, a stroke assessment system is provided and it comprises a medical data collection module **104** for receiving combinations of one or more of three different types of data, which are structured data element **101**, textural data element **102** and image data element **103**. The data collection module **104** may save in a data storage device, database or hardware computing device with processors and memory. The stroke assessment system further comprises a number of machine learning module; which are mortality prediction module **105**, ICD-9 code prediction module **106**, treatment module **107** and LVO prediction module **108**. These modules are configured to receive the data from data collection module **104.**

[0031] Usually, the data collection module **104** is configured to receive the data from one or more data sources. The received data may be saved in the data storage device, database or hardware computing device as an unstructured data set. For the structured data element **101**, textural data element **102,** and image data element **103**, a unique key or ID is required to identify each individual record or patient’s information. This unique key or ID is also used to make sure the data can be merged and synchronized correctly. In general, the unique key or ID may include record key, episode key, and patient key to make sure each data merge is done with the correct records. The data merging may be one to one, one to many, many to one or many to many.

[0032] The data collection module **104** is configured for data pre-processing the data from unstructured data to structured data. The data collection module **104** is further configured for converting features into numerical representations or using one-hot-coding in processing the data; and executing table merging in the data.

[0033] In the various embodiments, structured data element **101** is about information of patients and the clinical record. For example, the information may contain gender, age, weight, height, date and time of patient admission, laboratory test data, diagnosis name or ID designed by the hospital, indication on the diagnosis, patient may have principal diagnosis or non-principal diagnosis etc., and provide to the data collection module **104** as unstructured data. In general, patient demographics information in the structured data element **101** may also include a key for each patient to recognize and merge with other modules.

[0034] In the various embodiments, textual data element **102** contains text reports such as radiology examination text report, supplementary amended report, clinical and discharge note or any related clinical report. The radiology examination report may include the body text and the standardized code of examination of the radiology examination report. The clinical and discharge note may include the dates of discharge and/or admission and the body text of the clinical documentation. For the clinical report, a sequence order number by reference datetime among all clinical result datasets for identifying the clinical report record is necessary to ensure its correctness. For merging the patient records, an episode key assigned to the patient upon attendance for the examination and a unique patient key assigned to each individual patient are required by the data collection module **104**.

[0035] In the various embodiments, image data element **103** contains image data for the patients fulfilling some criteria by hospital, for example the CT Brain for brain plain. In this case, the CT Brain may contain computed and compressed tomography images, wherein the image compression may be made under a lossy-compression ratio, which means the quality of the images may be affected. In general, a sequence order number which references time among all image data datasets for identifying the image test is necessary to ascertain the patients' examination attendances for the data collection module **104**.

[0036] The mortality prediction module **105** is configured to predict the mortality of stroke patients from input received from data collection module **104**. The structured data and textual data used in mortality prediction module **105** are received from data collection module **104**. The data in data pre-processing element **109** includes, but is not limited to, stroke patients' demographics, diagnosis, family medical history, laboratory results and radiology reports, all of which are then used as input in the machine learning and deep learning models **113**. The demographics data includes, but is not limited to, gender and age. Family medical history data may contain indications of new symptoms or illnesses in a family through mapping of International Classification of Primary Care (ICPC) Code or other standard codes such as the ICD-9 code. The laboratory result data includes the results from chemical pathology test, haematology test, immunology test, microbiology, test and virology test. Due to the different testing methods and equipment, the detection unit of the sample needs to be unified in the process of data summarizing and pre-processing. As for the radiology examination report, body text of the examination reports, standardized codes of examinations are included.

[0037] The machine learning and deep learning models **113** include, but are not limited to, one or more of random forest (RF) classifier, Adaptive Boosting (AdaBoost), Extremely randomized trees (ExtraTree) classifier, XGBoost classifier, and TabNET, which constitute the first layer of the prediction model of the machine learning and deep learning models **113**. The second layer of the prediction model of the machine learning and deep learning models **113** is an ensemble model, which uses the input data to the machine learning and deep learning models **113** and results from the first layer as its input, then generates and outputs the probability of mortality **117**. The output result may have different timescale of the mortality prediction such as with 30-day, 6-month, or longer.

[0038] The ICD-9 prediction module **106** is configured to predict the specific subtype of strokes for stroke patients from input received from data collection module **104**. The structured data used in ICD-9 Prediction module **106** are received from data collection module **104**. The data in data pre-processing element **110** includes, but is not limited to, stroke patients' demographics, diagnosis, family medicine and laboratory results, all of which are then used as input in the machine learning and deep learning models **114**. The demographics data may contain gender, age, weight, height, etc. The family medical history data may contain indications of new symptoms or illnesses in a family through mapping of International Classification of Primary Care (ICPC) Codes or other standard codes such as the ICD-9 code. The laboratory result data includes the results from chemical pathology test, haematology test, immunology test, microbiology, test, and virology test. Due to the different testing methods and equipment, the detection unit of the sample needs to be unified in the process of data summarizing and pre-processing.

[0039] The machine learning and deep learning model **114** is configured to communicate with the ICD-9 prediction model data pre-processing element **110**. The machine learning and deep learning models **114** include, but are not limited to, one or more of RF classifier, AdaBoost, ExtraTree classifier, XGBoost classifier, and DNN or CNN models. Applying different models and parameters is required to achieve the best performance and output to provide accurate stroke subtype classification **118.** The output result is the specific subtype of stroke which corresponds to a particular ICD-9 code.

[0040] The treatment decision module **107** is configured to generate medication and treatment recommendations for stroke patients from input received from the data collection module **104**. The structured data, textual data, and image data used in Treatment module **106** are received from data collection module **104**. The data in data pre-processing element **111** includes, but is not limited to, stroke patients' demographics, diagnosis, family medical history, laboratory results, CT / MRI images, drug use, and medical procedures, all of which are then used as input in the reinforcement learning model **115**. The demographics data may contain gender, age, weight, and height etc. The family medical history data may contain indications of new symptoms or illnesses in a family through mapping of ICPC Codes or other standard codes such as ICD-9 code. The laboratory result data includes the results from chemical pathology test, haematology test, immunology test, microbiology, test and virology test. Due to the different testing methods and equipment, the detection unit of the sample needs to be unified in the process of data summarizing and pre-processing. The CT Brain for brain plain are in the format of Digital Imaging and Communications in Medicine (DICOM) and may contain computed and compressed tomography images, wherein the image compression may be made under a lossy-compression ratio, which means that the quality of the images may be affected. The drug use and medical procedures should provide the item code defined by government or hospital, drug name, dosage, and strength of the item for the drug. Also, dispensing date, duration, and the route form of the drug intake by the patient should also be included in the data of the data pre-processing element **111**.

[0041] Reinforcement learning model **115** is based on the SVM model or other machine learning methods or deep learning models such as DNN and CNN. Different models and parameters may be applied to achieve the best performance on the medication and treatment recommendations **119**. The output medication and treatment recommendations **119** includes drug name, dosage, duration, surgical operations and other clinical treatments for each patient which enable the patient to obtain the best prognostic effect.

[0042] For the LVO prediction module **108**, the certain system can provide the probability that the patient is an LVO patient by receiving data from data collection module **104**. The structured data and image data used in LVO prediction module **108** are derived from data collection module **104**. The data in data pre-processing element **112** includes, but is not limited to, stroke patients' demographics, diagnosis and CT images, all of which are then used as input in the deep learning models **116**. The demographics data may contain gender, age, weight, and height etc. Family medical history data, may contain indications of new symptoms or illnesses in a family through mapping of International Classification of Primary Care (ICPC) Codes or other standard codes such as ICD-9 code. The CT Brain for brain plain may contain computed tomography images, the Image compression may follow a ratio leveraged to have lossy compression, which meaning the quality of the images may affected.

[0043] The deep learning models **116** receive input from the data pre-processing element **112**. The deep learning models **116** comprises at least a deep learning model based on CNN, and/or a three-dimensional CNN (3DCNN) that may be one or more of

Densely Connected Convolutional Networks (DenseNets), EfficientNets, Squeeze-and-Excitation Networks (SENets or SEResNets) and a Vision Transformer (ViT). The deep learning models **116** apply different models and parameters to achieve the best performance and output LVO patient's classification **120**. The output LVO patient's classification 120 is provided a classification such as large-artery atherosclerosis, cardio embolism, small-vessel occlusion, and others (which are strokes of other determined etiology and stokes of undermined etiology).

[0044] FIG. 2 illustrates the data sources that by the stroke assessment system. The data collection process **206** includes the basic data collection **201** of the patient, the admission and discharge data **202** of the patient, the diagnosis information **203** from the doctor, the examination result **204** from the biological examination, and the examination result **205** from the imaging examination, etc. There are many types of data collected in the data collection process **206**, including structured data, textual data, and image data. Among them, structured data can be integer numbers, float numbers, numbers in string format, or empty; textual data requires that the text language be English; image data requires that the data be CT scan images or MRI images of the brain. The data collected through the data collection process **206** is called medical data **104**, or raw data. The raw data **301** in FIG. 3, the raw data **401** in FIG. 4, and the Digital Imaging and Communications in Medicine **501** in FIG. 5 are all part of the data collected by the data collection process **206**.

[0045] FIG. 3 illustrates the pre-processing pipeline of structured data used in the stroke assessment system. Raw data **301** are the data received from data collection process **206**. In certain embodiments, it is configured to generate a processed dataset for the machine learning models and deep learning models **113**, **114**, **115** and **116**. The raw data have patients' information and three different types of data which are structured data, textural data, and image data. Of which, only the structured data is sent to first filter **302**. The first filter **302** is used to filter non-stroke patients as they are not within the scope of this system. Some patients are classified as suspected patients because they have stroke-like symptoms, but after a series of clinical examinations, the diagnosis is cancelled or diagnosed with other diseases. Only samples that have been principle diagnosed as a stroke can pass through the first filter **302**. The second filter **303** preprocesses the data to remove samples with excessive missing values as well as other irrelevant samples. For a given feature, a certain threshold of missing values will be applied to eliminate samples with a missing rate equal to or greater than the threshold. Since there are different objectives for each of the four modules, the models present in the different modules have different requirements for the usage of the structured data. The models of each module select data related to their purpose as relevant data, and irrelevant data are excluded in that particular element. The data imputation **304**, which receives data from the second filter **303**, performs the imputation of missing data. Imputation is the process of replacing missing data (including “NA”, “None”, “Not record” etc.) with substituted values. Imputation methods including min-max imputation, mean imputation, non-negative matrix factorization, regression imputation, stochastic imputation, or multiple imputation method are applied in the data imputation **304**. The data imputation **304** provides the imputed data to data normalization **305**. Normalization refers to the process of adjusting values measured on different scales to a notionally common scale. Continuity data for different data ranges are redistributed within the range from 0 to 1 by data normalization **305** according to the distribution characteristics of the data itself. The data normalization **305** provides the normalized data to feature selection **306**. The feature selection **306** is a function generator element which provides the optimize data to models. Redundant data are eliminated by different feature selection methods. Different feature selectors are applied to calculate the importance of each feature, and features are selected according to the importance ranking. The output of the feature selection **306** is called the processed structured data **307**.

[0046] FIG. 4 illustrates the pre-processing pipeline of textual data used in the stroke assessment system. This element is a component of the elements **109** and element **111** in FIG. 1. The raw data **401** is the textual data part collected by the data collection process **206**. The raw data **401** is to pass through 3 layers of filters. The first filter **402** is used to filter non-stroke patients, which are irrelevant within the scope of the stroke assessment system. Some patients are classified as suspected patients because they have stroke-like symptoms, but after a series of clinical examinations, the diagnosis is cancelled or are diagnosed with other diseases. Only samples that have been principle diagnosed as a stroke can pass through the first filter **402**. The data passing through the filter **402** comes to the filter **403** which eliminates irrelevant reports. In the text report of the medical imaging examination, there are error reports, non-final reports, or other reports with insufficient information for data processing, and these data cannot pass the filter **403**. In addition, because the patient may take multiple CT or MRI during the entire admission treatment period, it reflects the condition of the disease at different stages during the treatment process. For example, in element **109**, the mortality of a patient is output after the first round of examinations after admission to the hospital. Therefore, a textual data that can represent a confirmed stroke or the first effective imaging examination after a stroke is selected to be used in the follow-up study. However, in element **111**, since the aim is to predict the patient's treatment, one or more data are selected for use according to the patient's treatment stage. The third filter **404** is used to pre-process the selected text data. It can delete stop words that do not contribute to the meaning of the research report, and organize the format of the text report. The sample after three filters is called processed textual data **405**.

[0047] FIG. 5A is a schematic block diagram illustrating one embodiment of a data extracting module. The DICOM from element **501** is the input of the first filter **502**, which excludes the slides with specific meta information from DICOM, including non-brain exam body parts, RGB planes, missing Image Position or Image Orientation information, diagnosis and patient ID. The following image normalization **503** is used to normalize the image size. All images are standardized as squares, such as 512 pixels * 512 pixels pictures. The input of the second filter are the processed structured data **307** from the structured data pre-processing process **300**, and the output from images normalization **503**, which can screen out the samples that only have one of structured data and image data. The images labelling **505** is designed to select and provide labels to the image that represents a confirmed stroke or the first effective imaging examination after a stroke is selected, because of the multi-level (patient level, accession level, and episode level) of images. The output of the images labelling **505** is the processed image data **506**.

[0048] FIG. 6 illustrates the machine learning and deep learning models’ training, validating and testing pipeline for structured data. This element is a component of the element **114** in FIG. 1. After receiving data from model **307**, data are randomly divided into training set **602**, validation set **603** and testing set **604**. Usually, the ratio of training set **602** may be imbalanced. Imbalanced data is a challenge for predictive modelling since most of algorithms are designed to handle balanced data. The Synthetic Minority Over-sampling TEchnique (SMOTE) is applied in element **602**, **603** and **604** to handle the imbalanced data. AUROC, accuracy AUPRC, precision, recall, and f1-score are applied to measure the performance of models in element **603** and **604**. The models with good performance are stored into element **607**.

[0049] FIG. 7 illustrates the machine learning, deep learning and ensemble models' training, validating and testing pipeline for structured and textual data. This element is a component of the element **113** in FIG. 1. The processed structured data **307** is from the structured data pre-processing process **300**, and the processed textual data **405** is from the textual data pre-processing process **400**. Element **701** is a combination filter, which can screen out the samples that only have one of structured data and textual data. All the data passing through the filter element **701** are mixed data **703**, where the structured data and textual data are filtered structured data **702** and filtered textual data **704**, respectively. The filtered structured data needs to go through machine learning and deep learning pipelines. The data are first randomly divided into training set **705**, validation set **706** and testing set **707**. The training set **705** and the validation set **706** are used to train various machine learning and deep learning models, and then the testing set **707** is added for model verification **702**, and finally trained machine learning and deep learning models are obtained and saved into element **715**. As for the filtered textual data **704**, the data is first divided into training set **708**, validation set **709** and testing set **710**. The training set **708** and the validation set **709** are used to train various machine learning and deep learning models, and then the testing set **714** is added for model verification **710**, and finally trained deep learning models are obtained and saved into element **716**. The input data of ensemble models **717** is mixed data **703**. All the ensemble models have a structure of two or more layers, the first layer is the combination of trained models from element **715** and **716**, the ensemble methods include, but are not limited to, various voting models and stacking models.

[0050] FIG. 8 illustrates the deep learning model’s training, validating and testing pipeline for structured and image data. This element is a component of the element **116** in FIG. 1. The processed structured data **307** is from the structured data pre-processing process **300**, and the processed textual data **506** is from the textual data pre-processing process **500**. Element **801** is a combination filter, which can screen out the samples that have only either structured data or image data. The data passing through the filter **801** are divided into structured data training set **802** and image data training set **803**, respectively. The training sets **802** and **803** are mixed into concatenating data and become the input of the 3DCNN and ViT training **805**. A plurality of backbones is used and compared in the 3DCNN, including Densely Connected Convolutional Networks (DenseNets), EfficientNets, Squeeze-and-Excitation Networks (SENets or SEResNets). We first input the image data into the models and then concatenate the embeddings from the last fully connected layers with clinical data. The prediction outputs of the 3DCNN and ViT are ensembled **806**. The LVO classifier is tested using the testing dataset. To improve performance and reduce generalization errors of the LVO classifier, test-time augmentation (TTA) and prediction averaging are added during the inference **807**.

[0051] FIG. 9 illustrates the machine learning, deep learning and reinforcement learning models’ training, validating and testing pipeline for structured data, image data and textual data. This element is a component of the element **115** in FIG. 1. The processed structured data **307** is from the structured data pre-processing process **300**, the processed textual data **405** is from the textual data pre-processing process **400**, and the processed image data is from the image data pre-processing process **500**. Element **901** is a combination filter configured to screen out the samples having one of structured data, textual data, and image data. All the data passing through the filter element **901** are mixed data **902**, where the structured data, image data, and textual data are filtered structured data **903**, filtered image data **904**, and filtered textual data **905**, respectively. The filtered structured data is to pass through the machine learning and deep learning pipelines. The data is first divided into training set **906**, validation set **907** and testing set **908**. The training set **906** and the validation set **907** are used to train various machine learning and deep learning models, and then the testing set **908** is added for model verification **916**. Finally, the trained machine learning and deep learning models are obtained and saved into element **921**. As for the filtered image data **904**, the data is first divided into training set **909**, validation set **910** and testing set **911**. The training set **909** and the validation set **910** are used to train various machine learning and deep learning models, and then the testing set **911** is added for model verification **918**, and finally trained deep learning models are obtained and saved into element **922**. As for the filtered textual data **905**, the data is first divided into training set **912**, validation set **913** and testing set **914**. The training set **912** and the validation set **913** are used to train various machine learning and deep learning models, and then the testing set **914** is added for model verification **920**, and finally trained deep learning models are obtained and saved into element **923**. The input data of reinforcement learning models **924** are mixed data **902** and output data from trained models **921**, **922**, and **923**.

[0052] The foregoing description of the present invention has been provided for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise forms disclosed. Many modifications and variations will be apparent to the practitioner skilled in the art.

[0053] The embodiments were chosen and described in order to best explain the principles of the invention and its practical application, thereby enabling others skilled in the art to understand the invention for various embodiments and with various modifications that are suited to the particular use contemplated.

Claims:

What is claimed is:

1. A machine learning and deep learning-based stroke assessment system, comprising:

a clinical data collection module configured to extract stroke patients related clinical data from one or more data sources, organize the extracted stroke patients related clinical data into one or more of structured, textual, and image datasets; and preserve patient identity information or encrypted identity information in the clinical data records to ensure that the clinical data records of same patient are having same patient identifier and the clinical data records of same episode are having same episode key;

a machine learning and deep learning-based mortality prediction module configured to analyse the structured datasets and the textual datasets to predict a long- or short-term mortality rate of the stroke patients referred to in the extracted stroke patients related clinical data;

a machine learning-based accurate stroke subtype classification module configured to analyse the structured datasets to generate a classification of stroke subtypes suffered by the stroke patients referred to in the extracted stroke patients related clinical data;

a machine learning and deep learning-based treatment decision module configured to analyse the structured datasets, the textual datasets, and the image datasets to generate medical decisions for treating the stroke patients; and

a deep learning-based large vessel occlusion (LVO) patient classification module configured to analyse the structured datasets and the image datasets to screen out LVO patients from ischemic stroke patients;

wherein each of the modules comprises one or more computer processors specially configured by one or more machine instructions stored in one or more non-transient memory circuities.

2. The system of claim 1, wherein each of the modules comprises one or more supervised machine learning and deep learning models configured specifically to process and analyse one or more of the structured, textual, and image datasets organized from the clinical data.

3. The system of claim 1,

wherein the mortality prediction module comprises one or more ensemble models; and

wherein the ensemble models are logical multi-level structures in which one or more higher-level models are configured to process input data to the mortality prediction module and output data from one or more lower-level models in the logical multi-level structures.

4. The system of claim 1,

wherein the treatment decision module comprises one or more reinforcement learning models; and

wherein the reinforcement learning models are logical multi-level structures in which one or more higher-level models are configured to process input data to the treatment decision module and output data from one or more lower-level models in the logical multi-level structures.

5. The system of claim 1, further comprising:

a first filter configured to select from the structured dataset data records of patients classified as stroke patients as principal diagnosis data records; and eliminate from the structured dataset data non-principal diagnosis data records;

a second filter configured to eliminate from the structured dataset data records with excessive missing values, and/or of irrelevant samples; wherein sample relevancy is based on selected demographics, diagnosis, family history, and laboratory results; and wherein the data records with excessive missing values are data records having a number of missing values larger than a defined threshold for number of missing values;

a data imputation module configured to apply one or more of min-max imputation method, mean imputation method, non-negative matrix factorization method, regression imputation method, stochastic imputation method, and multiple imputation method to replace the missing values in the structured dataset data records; and

a data normalization module configured to adjust one or more values in one or more of the structured datasets data records to within a range from 0 to 1.

6. The system of claim 1, further comprising:

a first filter configured to select from the textual dataset reports of patients classified as stroke patients as principal diagnosis reports; and eliminate from the textual dataset non-principal diagnosis reports;

a second filter configured to eliminate from the textual dataset irrelevant reports including error reports, non-final reports, and reports that are unrelated to the principal diagnosis reports; and

a third filter configured to delete stop words that do not contribute to the meaning of the research report, and organizes the format of the text report.

7. The system of claim 1, further comprising:

a first filter configured to eliminate from the image dataset slides with specific meta information from DICOM, the specific meta information includes one or more of non-brain exam body parts, RGB planes, missing or corrupted image position, missing or corrupted image orientation information, missing or corrupted diagnosis, and missing or corrupted patient ID;

an image normalization module configured to standardize all the images as squares;

the second filter eliminates excessive missing values samples and irrelevant samples; the images labelling element assigns labels to the selected images representing the patient’s diagnosis.

8. The system of claim 2,

wherein the machine learning models that are specifically configured to process and analyse the structured dataset comprise:

a random forest (RF) classifier,

an Adaptive Boosting (AdaBoost),

an Extremely randomized trees (ExtraTree) classifier, and

a XGBoost classifier;

wherein the deep learning models that are specifically configured to process and analyse the structured dataset comprise a TabNet classifier;

wherein during training, validation, and testing of the machine learning models and the deep learning models, random sampling is applied to the structured dataset, dividing the structured dataset into one or more of training datasets, validation datasets, and testing datasets;

wherein each of the machine learning models and the deep learning models is trained respectively using the training dataset;

wherein each of the machine learning models and the deep learning models is evaluated by using AUC value of the validation set, and the models with highest AUC values are selected as the best performing machine learning models and deep learning models for run-time; and

wherein during the run-time, the structured dataset is reloaded to the best performing machine learning models and deep learning models to generate one or more prediction results.

9. The system of claim 2,

wherein the deep learning models that are specifically configured to process and analyse the textual dataset comprise a DistilBERT;

wherein during training, validation, and testing of the deep learning models, random sampling is applied to the textual dataset, dividing the structured dataset into one or more of training datasets, validation datasets, and testing datasets;

wherein the deep learning models are loaded on to multiple-block GPUs and trained in parallel using the training dataset;

wherein each of the deep learning models is evaluated by using AUC value of the validation set, and the models with highest AUC values are selected as the best performing deep learning models for run-time; and

wherein during the run-time, the textual dataset is reloaded to the best performing deep learning models to generate one or more prediction results.

10. The system of claim 2,

wherein the deep learning models that are specifically configured to process and analyse the image dataset comprise:

a three-dimensional convoluted neural network (3DCNN), and

a Vision Transformer (ViT);

wherein during training, validation, and testing of the deep learning models, random sampling is applied to the textual dataset, dividing the structured dataset into one or more of training datasets, validation datasets, and testing datasets;

wherein the deep learning models are loaded on to multiple-block GPUs and trained in parallel using the training dataset;

wherein test-time augmentation (TTA) is applied to the testing dataset and test prediction results are averaged during inference during the testing of the deep learning models;

wherein each of the deep learning models is evaluated by using AUC value of the validation set, and the models with highest AUC values are selected as the best performing deep learning models for run-time; and

wherein during the run-time, the image dataset is reloaded to the best performing deep learning models to generate one or more prediction results.

11. The system of claim 3,

wherein the ensemble models are one or more of voting ensemble models and stacking ensemble models;

wherein validation results from structured dataset and textual dataset are merged according to the patient information and formed as input to the ensemble models;

wherein a weight of each result from each of the ensemble models is decided by one or more loss values during the training and testing; and

wherein a logistic regression is applied as a secondary model for stacking ensemble model to avoid overfitting.

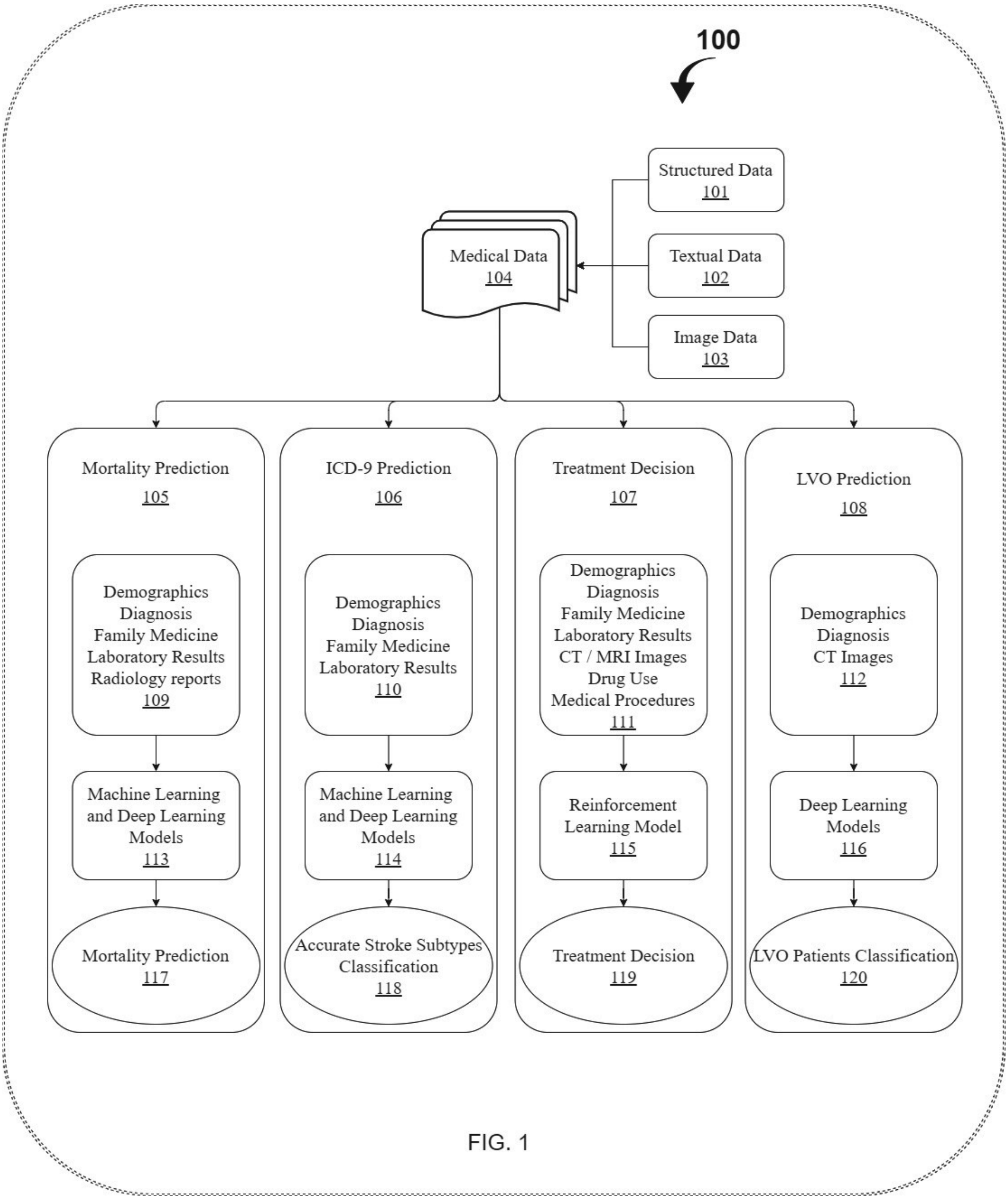
100
Structured Data
101
Textual Data
102
Image Data
103
Medical Data
104
Mortality Prediction
105
ICD-9 Prediction
106
Treatment Decision
107
LVO Prediction
108
Demographics
Diagnosis
Family Medicine
Laboratory Results
Radiology reports
109
Demographics
Diagnosis
Family Medicine
Laboratory Results
110
Demographics
Diagnosis
Family Medicine
Laboratory Results
CT / MRI Images
Drug Use
Medical Procedures
111
Demographics
Diagnosis
CT Images
112
Machine Learning
and Deep Learning
Models
113
Machine Learning
and Deep Learning
Models
114
Reinforcement
Learning Model
115
Deep Learning
Models
116
Mortality Prediction
117
Accurate Stroke Subtypes
Classification
118
Treatment Decision
119
LVO Patients Classification
120

FIG. 1

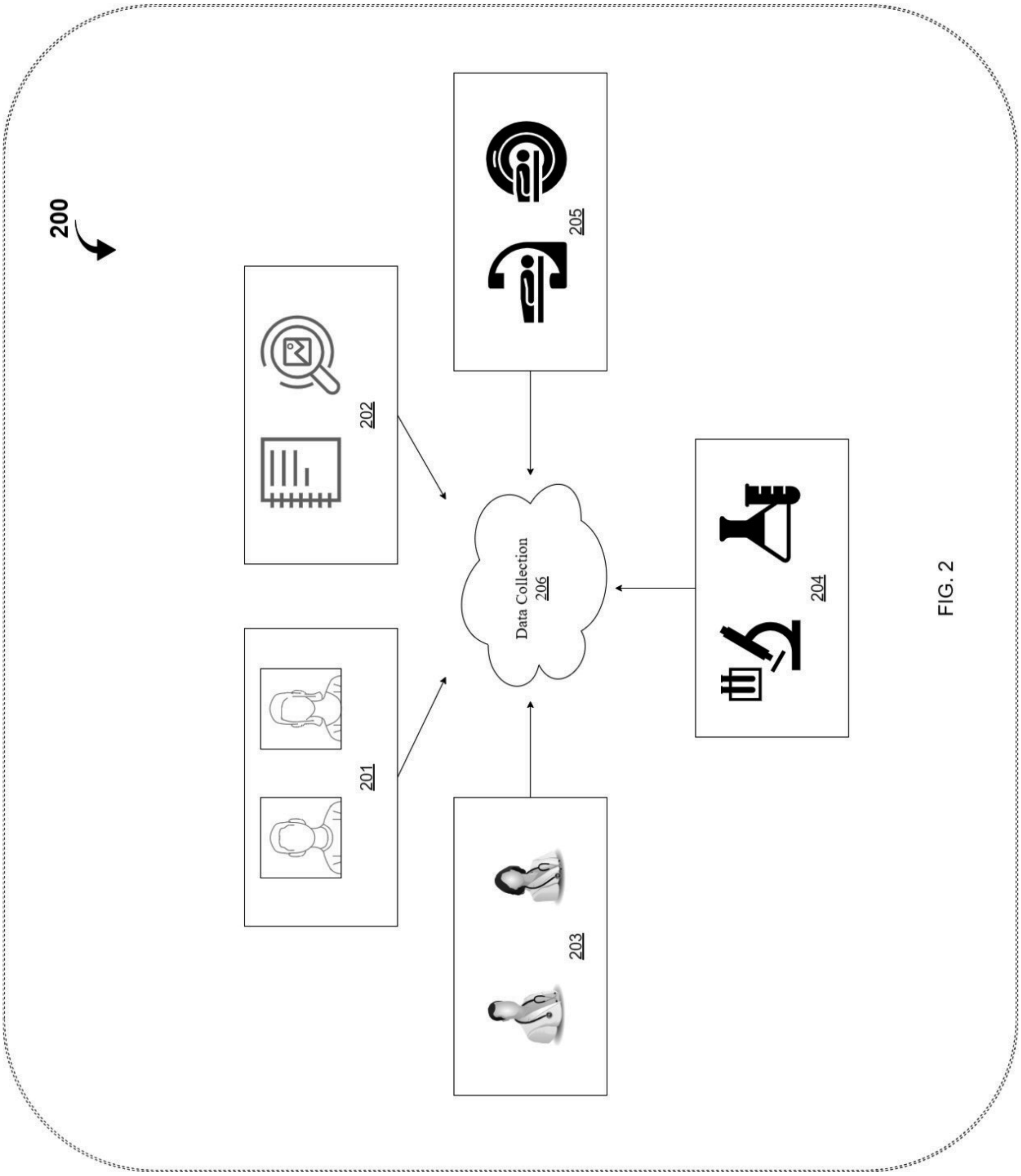

FIG. 2

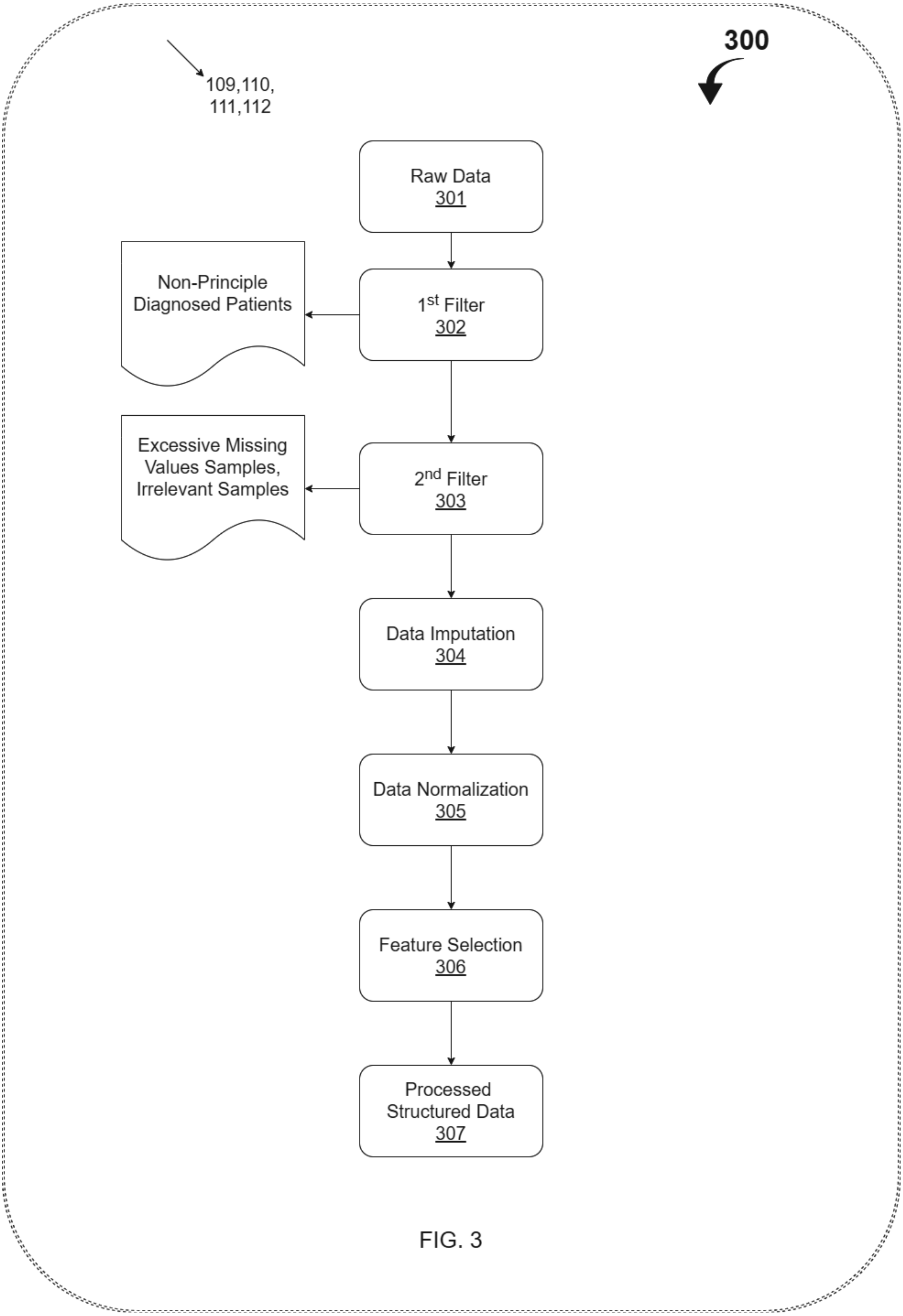
300
109,110,
111,112
Raw Data
301
1st Filter
302
Non-Principle
Diagnosed Patients
2nd Filter
303
Excessive Missing
Values Samples,
Irrelevant Samples
Data Imputation
304
Data Normalization
305
Feature Selection
306
Processed
Structured Data
307

FIG. 3

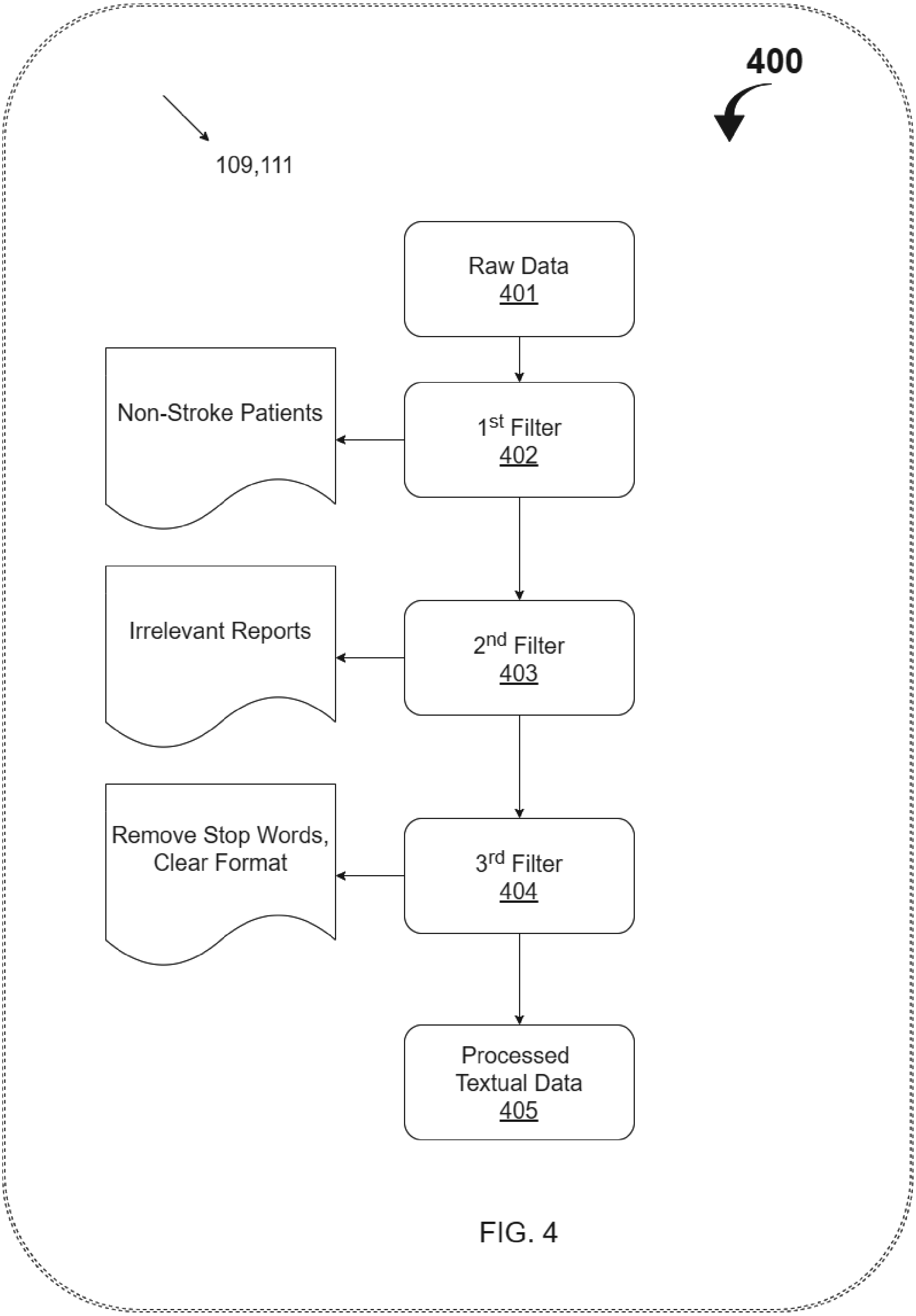
400
109,111
Raw Data
401
1st Filter
402
Non-Stroke Patients
2nd Filter
403
Irrelevant Reports
3rd Filter
404
Remove Stop Words, Clear Format
Processed Textual Data
405

FIG. 4

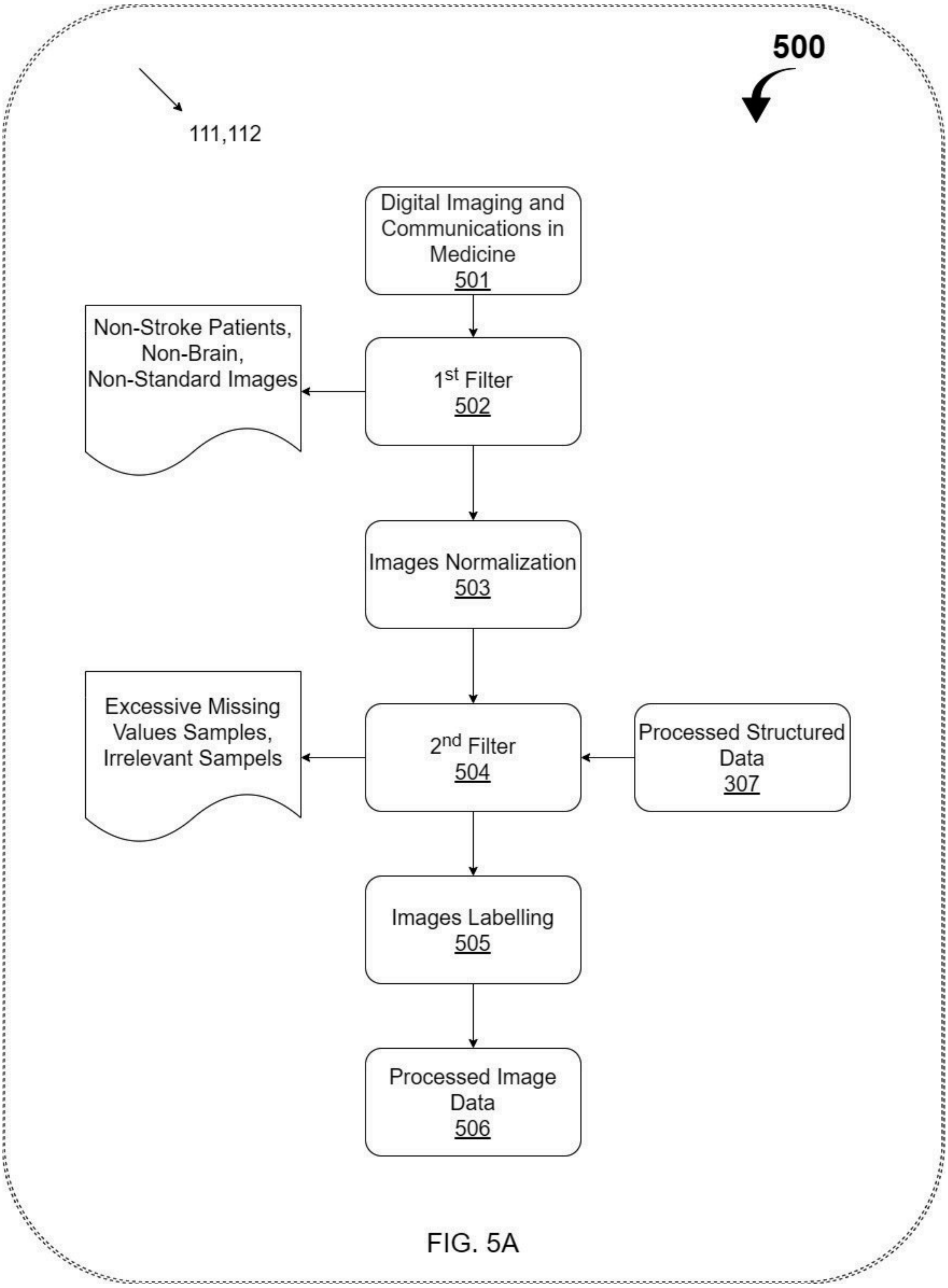
500
111,112
Digital Imaging and Communications in Medicine
501
Non-Stroke Patients, Non-Brain, Non-Standard Images
1st Filter
502
Images Normalization
503
Excessive Missing Values Samples, Irrelevant Sampels
2nd Filter
504
Processed Structured Data
307
Images Labelling
505
Processed Image Data
506

FIG. 5A

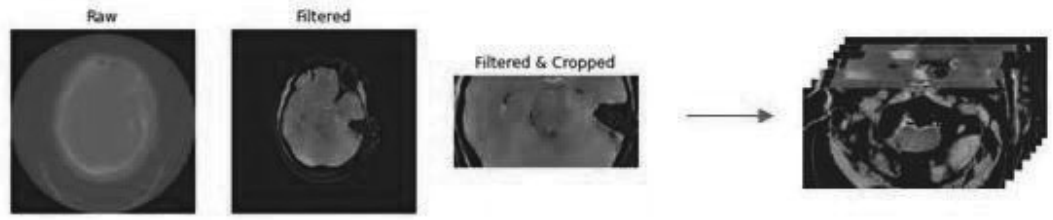

FIG. 5B

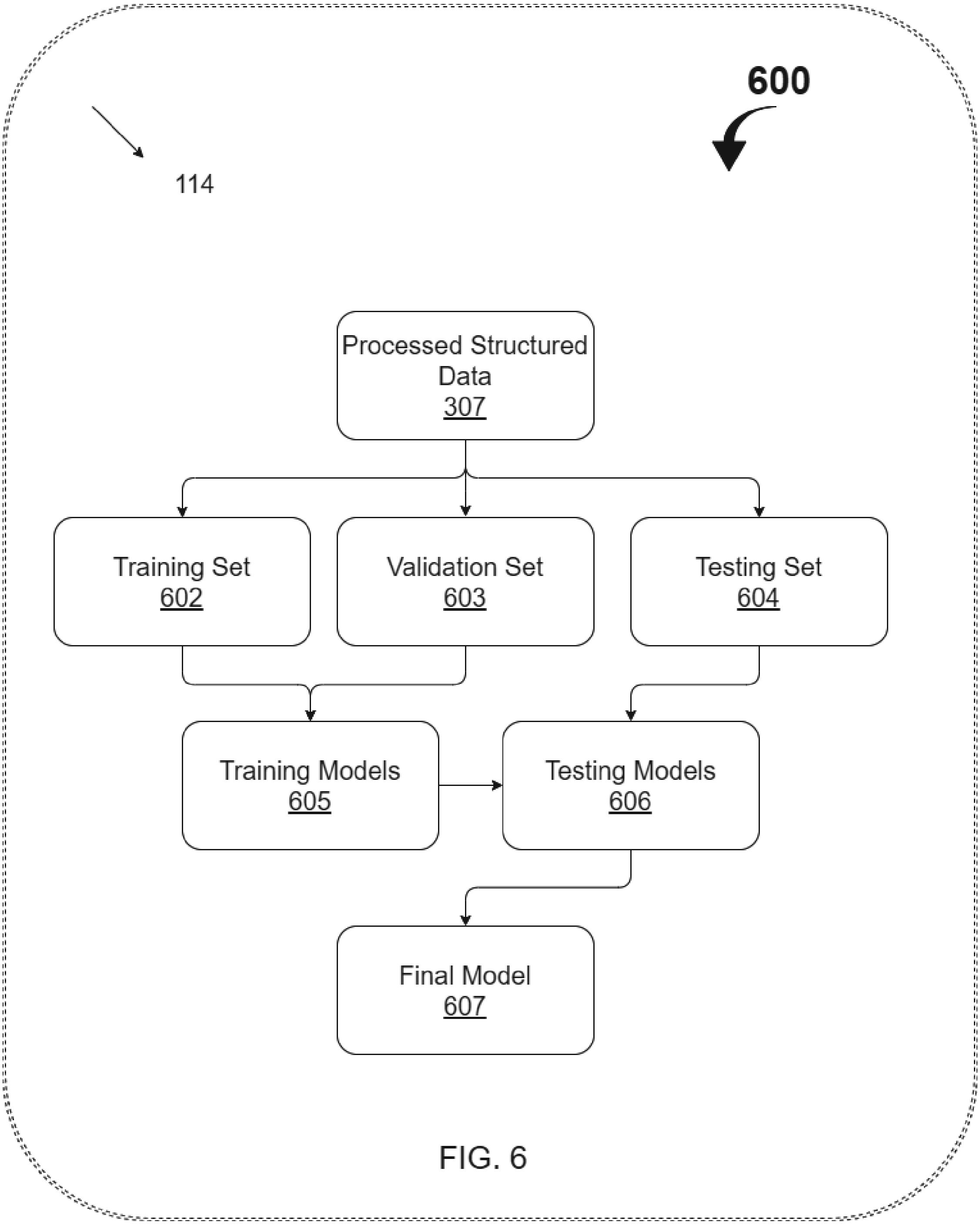
600
114
Processed Structured Data 307
Training Set 602
Validation Set 603
Testing Set 604
Training Models 605
Testing Models 606
Final Model 607

FIG. 6

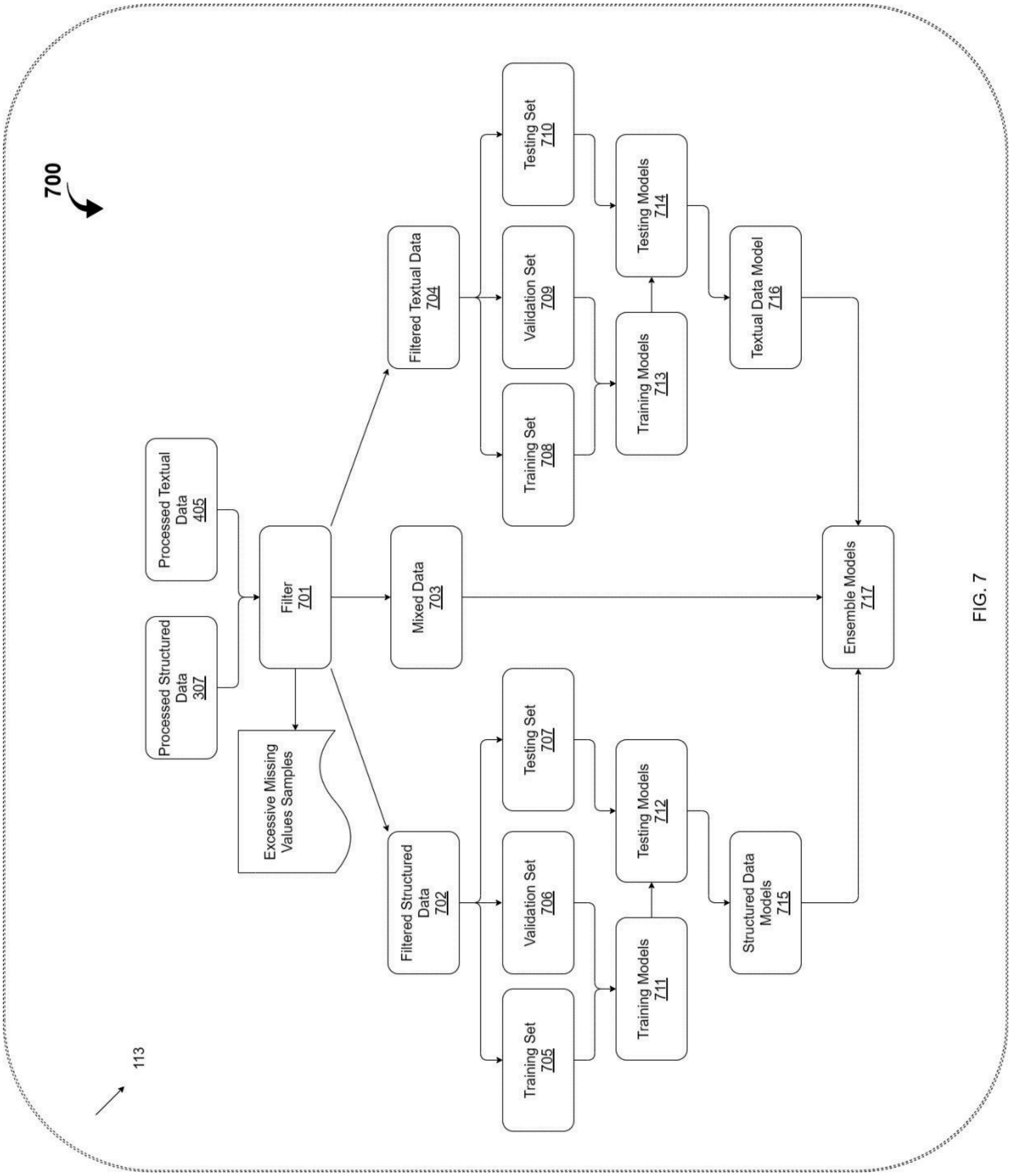
700
113
Processed Structured Data 307
Processed Textual Data 405
Filter 701
Excessive Missing Values Samples
Filtered Structured Data 702
Mixed Data 703
Filtered Textual Data 704
Training Set 705
Validation Set 706
Testing Set 707
Training Set 708
Validation Set 709
Testing Set 710
Training Models 711
Testing Models 712
Training Models 713
Testing Models 714
Structured Data Models 715
Textual Data Model 716
Ensemble Models 717

FIG. 7

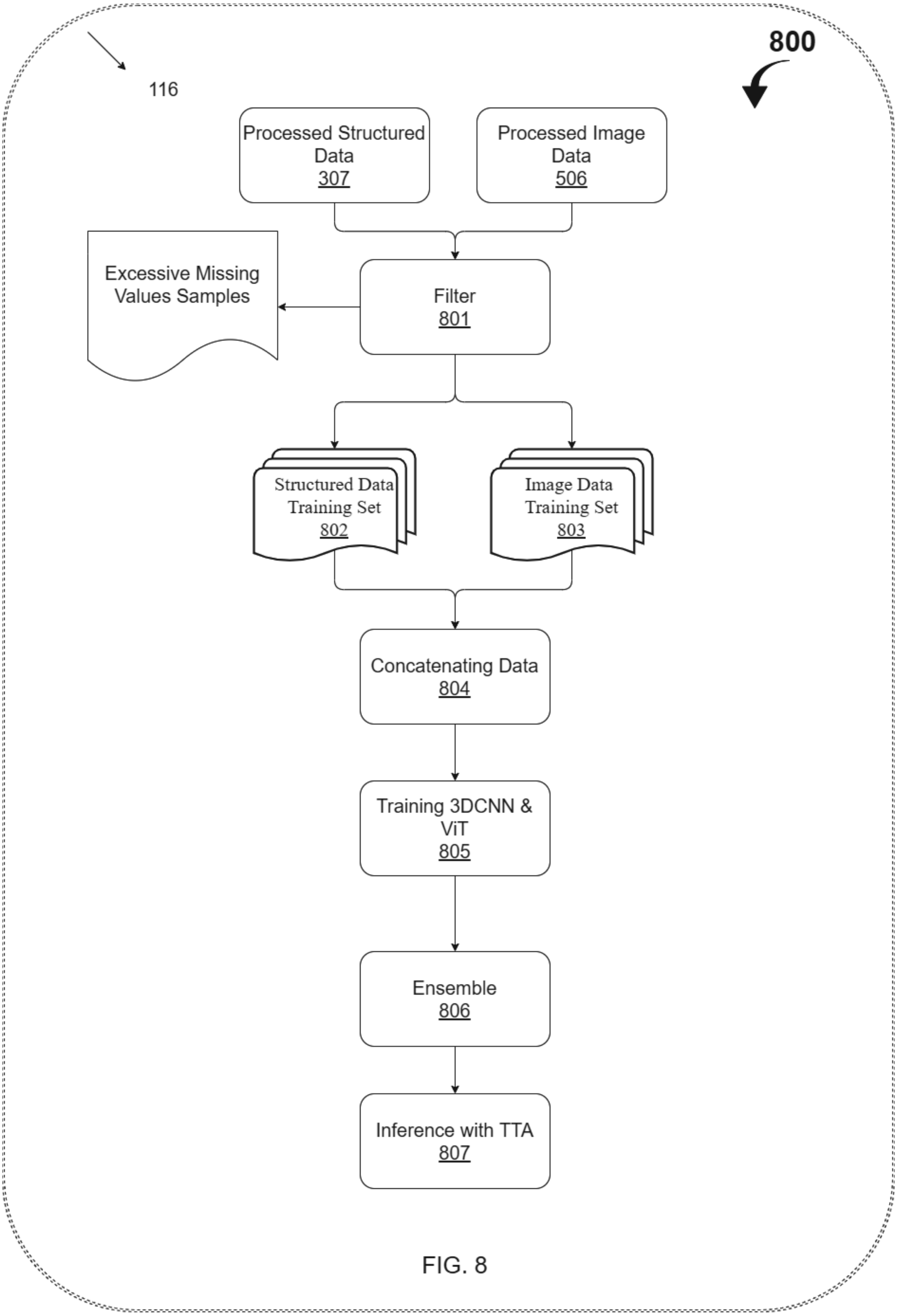
116
800
Processed Structured Data 307
Processed Image Data 506
Excessive Missing Values Samples
Filter 801
Structured Data Training Set 802
Image Data Training Set 803
Concatenating Data 804
Training 3DCNN & ViT 805
Ensemble 806
Inference with TTA 807

FIG. 8

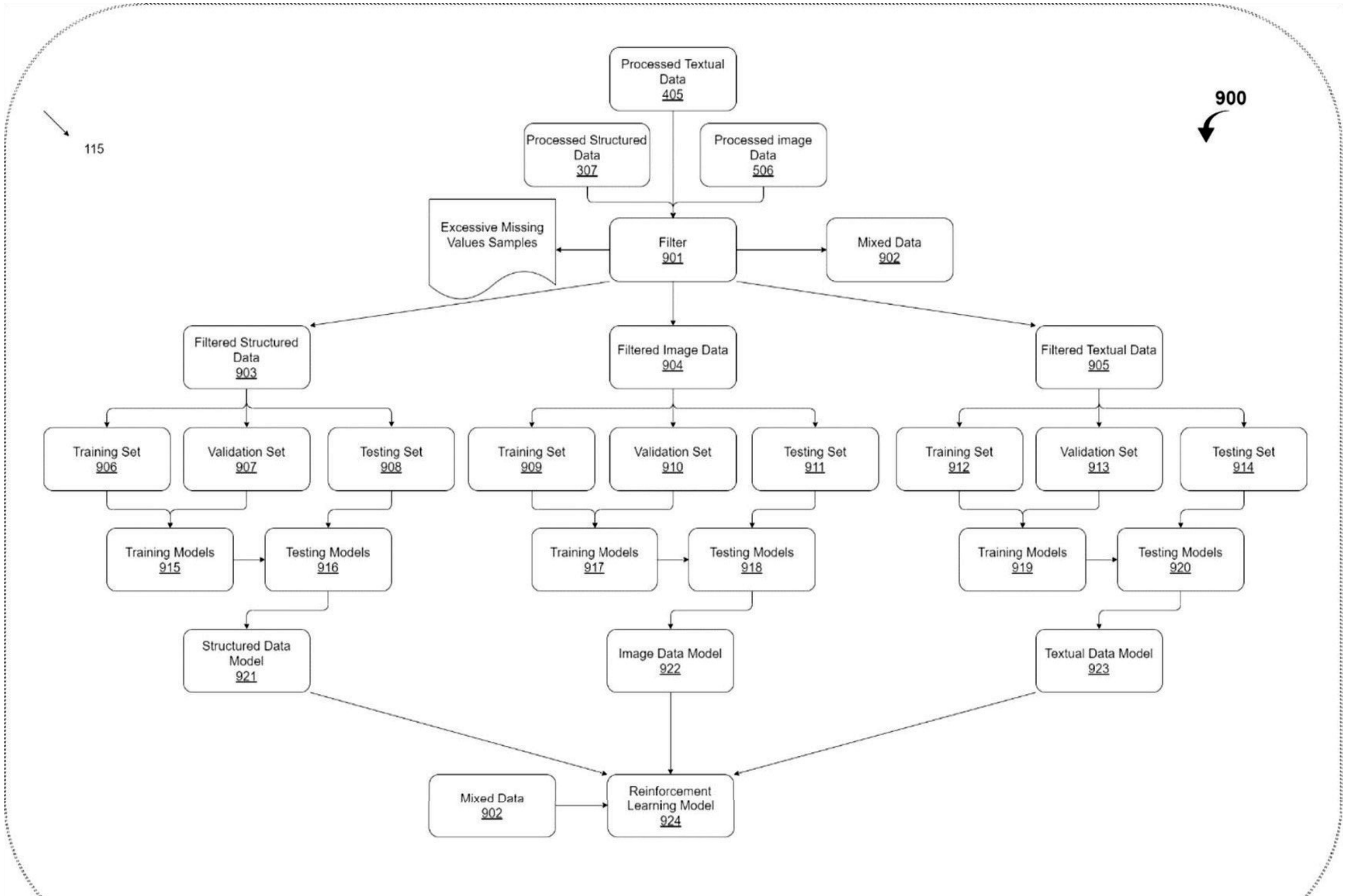
900
115
Processed Textual Data 405
Processed Structured Data 307
Processed image Data 506
Excessive Missing Values Samples
Filter 901
Mixed Data 902
Filtered Structured Data 903
Filtered Image Data 904
Filtered Textual Data 905
Training Set 906
Validation Set 907
Testing Set 908
Training Set 909
Validation Set 910
Testing Set 911
Training Set 912
Validation Set 913
Testing Set 914
Training Models 915
Testing Models 916
Training Models 917
Testing Models 918
Training Models 919
Testing Models 920
Structured Data Model 921
Image Data Model 922
Textual Data Model 923
Mixed Data 902
Reinforcement Learning Model 924

FIG. 9