**Table 12:** Common datasets used for validation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No of FS** | **Tittle** | **Name of dataset** | **Description of dataset** | **Size of dataset** |
| [FS1] | Shifu: Deep Learning Based Advisor-advisee Relationship Mining in Scholarly Big Data. | PhDTree DBLP dataset. | A deep-learning-based advisor-advisee relationship identification method which takes into account both the local properties and network characteristics | Consist of Inferred 1,111,513 advisor- advisee pairs |
| [FS2] | GPU-Accelerated Parallel Hierarchical Extreme Learning Machine on Flink for Big Data | MINST,GISETTE, ADULT, WINE | Dataset of Flink cluster | Average running time of large-scale training samples with 2,000,000 |
| [FS3] | Learning deep representation with large-scale attributes | ILSVRC 2014 | large-scale object attribute database | Contains rich attribute annotations (over 300 attributes) for ∼180k samples and 494 object classes |
| [FS4] | Large-scale deep learning for computer-aided detection of mammographic lesions | mammographic data set | Dataset that consist of an X-rays with diagnosed tumors of the breasts | The train and validation set comprise 44,090 mammographic views, from which we used 39,872 for training and 4218 for validation |
| [FS5] | Mining Fashion Outfit Composition Using an End-to-End Deep Learning Approach on Set Data | Outfit dataset | large-scale fashion outfit dataset with | 195K outfits and 368K fashion items |
| [FS6] | A novel sparse representation classification face recognition based on deep learning | FERET face database, | Database of a face images | 1400 images, including a total of 200 different people (seven of each face image) |
| [FS7] | T-LRA: Trend-Based Learning Rate Annealing for Deep Neural Networks | Multimedia dataset | Multimedia dataset  with different images | This dataset includes 60,000 images (50,000 for training and 10,000 for testing) |
| [FS8] | Adaptive neuron apoptosis for accelerating deep learning on large scale systems | Higgs Boson data set | Not specified | Classification dataset (11M samples) and ImageNet classification datasets (≈1.3M images) |
| [FS9] | Retrieval from and Understanding of Large-Scale Multimodal Medical Datasets : A Review | Clinical data | Medical datasets in general. Some are Generated through xray | (small: < 1000 data items, medium, or large: > 10000), 3D microscopy images (17,107 images), 600 images in 2004 and 300,000 images in 2013 |
| [FS10] | Big Data and Deep Analytics Applied to the Common Tactical Air Picture (CTAP) and Combat Identification (CID) | (NVESD) | Visible and IR imagery collected by the US Army Night Vision and Electronic Sensors Directorate | 207 GB of IR imagery and 106 GB of visible imagery 4500 total images per test |
| [FS11] | Mobile Big Data Analytics Using Deep Learning and Apache Spark | Actitracker dataset | Which includes accelerometer samples of 6 conventional activities (walking, jogging, climbing stairs, sitting, standing, and lying down) from 563 crowdsourcing users | Both labeled and unlabeled data of 2,980,765 and 38,209,772 samples, respectively |
| [FS12] | Comparison between Multi-Class Classifiers and Deep Learning with Focus on Industry 4.0 | H2O cluster | datasets to the H2O cluster | First 16,000 items from dataset. Other 4,000 items from dataset |
| [FS13] | Deep Neural Networks for Traffic Flow Prediction | The traffic flow condition data | The traffic flow condition data | The central server produces 6,260,603 traffic condition datasets every five minutes, 288 times per day. That is, the traffic flow condition data consists of 180,305,364 values (6,260,603-by-288 matrix) for each day |
| [FS14] | Gender Classification by Deep Learning on Millions of Weakly Labelled Images | Face images  dataset | Dataset of face images | Five million weakly labeled face images |
| [FS15] | Large Deep Neural Networks for MS Lesion Segmentation | (CLIMB) | The Partners Multiple Sclerosis center enrolled in the CLIMB | 3000 patients enrolled in the CLIMB study |
| [FS16] | Deep net architectures for visual-based clothing image recognition on large database | Clothing datasets | Clothing dataset | Dataset with 80,000 images |
| [FS17] | Deep Computation Model for Unsupervised Feature Learning on Big Data | CUAVE Dataset, SNAE2 Dataset, INEX 2007 Dataset | Not specified | 500 training images, 800 testing images 100,000 unlabeled images 1,800 video clips grouped into four categories |
| [FS18] | Social Network Analysis of TV Drama Characters via Deep Concept Hierarchies | Data of TV drama | Data of TV drama | Adopting approximately 4400-minute data of TV drama |
| [FS19] | Small boxes big data: A deep learning approach to optimize variable sized bin packing | Real-world customer produced logistics orders of Walmart e-Commerce. | Real-world customer produced logistics orders of Walmart e-Commerce | 4 million data, 4,278,645 instances |
| [FS20] | Big-Data-Generated Traffic Flow Prediction Using Deep Learning and Dempster-Shafer Theory | (PeMS), CityPulse Dataset | Traffic flow datasets.  Caltrans Performance Measurements Systems (PeMS), CityPulse Dataset | There are 47 roads, where the traffic flow of each road is calculated based on the average of all the loop detectors in that particular road |
| [FS21] | Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework | SAT-4 and SAT-6 | Were extracted from the NASA National Agriculture Imagery Program dataset | SAT-4 consists of 5,00,000 images, SAT-6 consists of 4,05,000 images |
| [FS22] | DP-miRNA: An improved prediction of precursor microRNA using deep learning model | pre-miRNA | The human pre-miRNA sequence | The negative dataset consists of 8494 pseudo hairpins |
| [FS23] | Predicting the Success of Bank Telemarketing using Deep Convolutional Neural Network | Bank marketing data | Bank marketing data | 45,211 instances of whether acceptance or rejection to the phone call proposal for the given deposit option are collected in 2008~2010, by the Portuguese banking institution |
| [FS24] | A Deep Learning Approach to Android Malware Feature Learning and Detection | Benign apps | Not specified | In an experiment with 3,986 benign apps and 3,986 malware |
| [FS25] | Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography | Mammographic  dataset | Mammographic  dataset | After data augmentation, a total of 45,072 mammographic ROIs and 37,450 DBT ROIs were obtained |
| [FS26] | A Novel Multimode Fault Classification Method Based on Deep Learning | Case Western Reserve University Bearing | Case Western Reserve University Bearing | This paper selects 200 samples in each fault type; each sample contains 2048 observation points. 100 samples are randomly selected as the training data, and the other 100 samples as the testing data |
| [FS27] | Learning Transportation Modes from Smartphone Sensors Based on Deep Neural Network | Accelerometer, magnetometer, and gyroscope measurements database | Accelerometer, magnetometer, and gyroscope measurements data base | The proposed mechanism is evaluated on a database that contains more than 1000 h of accelerometer, magnetometer, and gyroscope measurements from five transportation modes, including still, walk, run, bike, and vehicle |
| [FS28] | Automated IT system failure prediction: A deep learning approach | (WSC) and (MSC) | The dataset has been collected from two large enterprise systems, a web server cluster (WSC) and a mailer server, cluster (MSC) | (WSC) 1,885,022 (81.4%) (MSC) 4,536,360 (96.7%) |
| [FS29] | Deep Convolutional Computation Model for Feature Learning on Big Data in Internet of Things | CUAVE, SNAE2, and, STL-10 | The SNAE2 is collected from YouTube | It contains more than 1800 pieces, whose theme focuses on sport, news, advertisement and entertainment. The STL-10 dataset contains more than 1300 images, where 500 images |
| [FS30] | Fast auto-clean CNN model for online prediction of food materials | (MLC dataset) | The dataset is collected by a large food supply chain platform in China (www.mealcome.com), it includes nearly | 1000 restaurants and above 12 000 food supplies |
| [FS31] | Large-scale restricted Boltzmann machines on single GPU | RBM, regCuCD-1, cuCD-1 | Not specified | No of parameters 32M 64M 128M 256M 512M 1B 2B |
| [FS32] | Weakly Semi-supervised Deep Learning for Multi-label Image Annotation | NUS-WIDE, MS COCO 2014 | Not specified | Labeled images 82; 783 74; 320 64; 430 48; 265 |
| [FS33] | A Novel Left Ventricular Volumes Prediction Method Based on Deep Learning Network in Cardiac MRI | ADSB datasets | Direct LV volumes prediction research using the new open accessible ADSB datasets | This datasets include 1140 subjects (more than 1026000 CMR images) |