Supplementary Information
Improving Hospital Length of Stay Prediction through Heterogeneous Data Integration from MIMIC-III Records

Ahmad F. Al Musawi, Pratip Rana, Sibtanu Raha, William C. Sleeman IV, Rishabh Kapoor, and Preetam Ghosh

1 Structured Data

1.1 Diagnoses

Here is the list of diagnoses ICD-9 codes (first three digits) used in this study. {008, 038, 041, 042, 070, 112, 162, 196, 197, 198, 211, 238, 244, 250, 253, 255, 263, 272, 274, 275, 276, 278, 280, 284, 285, 286, 287, 288, 289, 291, 292, 293, 294, 295, 296, 300, 303, 304, 305, 311, 327, 331, 333, 338, 342, 344, 345, 346, 348, 349, 357, 362, 365, 396, 397, 398, 401, 403, 410, 411, 412, 413, 414, 415, 416, 423, 424, 425, 426, 427, 428, 429, 430, 431, 433, 434, 437, 438, 440, 441, 443, 453, 455, 456, 458, 459, 478, 482, 486, 491, 492, 493, 496, 507, 511, 512, 518, 519, 530, 531, 532, 535, 536, 537, 553, 557, 560, 562, 564, 567, 568, 569, 570, 571, 572, 574, 576, 577, 578, 583, 584, 585, 593, 599, 600, 682, 693, 707, 710, 714, 715, 719, 721, 724, 728, 729, 730, 733, 745, 746, 747, 765, 769, 770, 771, 774, 775, 776, 779, 780, 781, 782, 784, 785, 786, 787, 788, 789, 790, 799, 801, 802, 805, 807, 808, 852, 860, 861, 873, 995, 996, 997, 998, 999, E81, E84, E87, E88, E91, E92, E93, E94, E95, V05, V09, V10, V12, V15, V17, V29, V30, V31, V42, V43, V44, V45, V46, V49, V50, V58, V64, V70, V85}.

1.2 Procedures

Here is the list of procedures ICD-9 codes (first two digits) used in this study. {10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 91, 92, 93, 94, 95, 96, 97, 98, 99}

1.3 Medications

Here is the list of final processed medications extracted from MIMIC-III PRESCRIPTIONS table. {*NF* Ertapenem Sodium, *NF* Rasburicase, 0.45% Sodium Chloride, 0.9% Sodium Chloride, 0.9% Sodium Chloride (Mini Bag Plus), 1/2 NS, 5% Dextrose, AMP, AcetaZOLamide Sodium, Acetaminophen, Acetaminophen (Liquid), Acetaminophen IV, Acetaminophen w/Codeine, Acetylcysteine 20%, Acular, Acyclovir, Adenosine, Albumin 25% (12.5g / 50mL), Albumin 5% (12.5g / 250mL), Albuterol, Albuterol 0.083% Neb Soln, Albuterol Inhaler, Albuterol MDI, Albuterol-Ipratropium, Alendronate Sodium, Allopurinol, Alprazolam, Alteplase (Catheter Clearance), Aluminum Hydroxide Suspension, Aluminum-Magnesium Hydrox.-Simethicone, Amino Acids 4.25% W/ Dextrose 5%, Amiodarone, Amiodarone HCl, Amlodipine, Amphotericin B, Ampicillin Sodium, Ampicillin-Sulbactam, Artificial Tear Ointment, Artificial Tears, Artificial Tears Preserv. Free, Ascorbic Acid, Ascorbic Acid (Liquid), Aspirin, Aspirin EC, Atenolol, Atorvastatin, Atovaquone Suspension, Atropine Sulfate, Azithromycin, Azithromycin , Bacitracin/Polymyxin B Sulfate Opht. Oint, Baclofen, Bag, Benzonatate, Bisacodyl, Brimonidine Tartrate 0.15% Ophth., Bupivacaine 0.1%, Bupivacaine 0.75%, Butalbital-Acet-Caffeine, Calcitriol, Calcium Carbonate, Calcium Chloride, Calcium Gluconate, Calcium Replacement (Oncology), Caphosol, Capsaicin 0.025%, Captopril, Carbamazepine XR, Carbidopa-Levodopa (10-100), Carbidopa-Levodopa (25-250), Carvedilol, CefTRIAXone, CefazoLIN, Cefazolin, CefePIME, Cefpodoxime Proxetil, CeftazIDIME, Ceftazidime, CeftriaXONE, Ceftriaxone, Cepacol (Menthol), Cephalexin, Cetylpyridinium Chl (Cepacol), Chlorhexidine Gluconate 0.12% Oral Rinse, Cilostazol, Ciprofloxacin, Ciprofloxacin 0.3% Ophth Soln, Ciprofloxacin HCl, Ciprofloxacin IV, Cisatracurium Besylate, Citalopram, Citalopram Hydrobromide, Citrate Dextrose 3% (ACD-A) CRRT, Clindamycin, Clindamycin Suspension, Clonazepam, CloniDINE, Clonidine HCl, Clonidine Patch 0.1 mg/24 hr, Clonidine Patch 0.3 mg/24 hr, Clopidogrel, Clopidogrel Bisulfate, Clotrimazole, Codeine Sulfate, Collagenase Ointment, Cosyntropin, Creon 12, Cromolyn, Cyanocobalamin, Cyclophosphamide, Cyproheptadine, Cytarabine, D5 1/2NS, D5LR, D5NS, D5W, D5W (EXCEL BAG), DOBUTamine, DOPamine, DOXOrubicin, Dapsone, Daptomycin, Demeclocycline, Desitin, Dexamethasone, Dexamethasone Sod Phosphate, Dextrose 5%, Dextrose 50%, Diazepam, Dicloxacillin, Digoxin, Diltiazem, Diltiazem Extended-Release, DiphenhydrAMINE, Diphenhydramine HCl, Docusate Sodium, Docusate Sodium (Liquid), Dolasetron Mesylate, Donepezil, Donepezil 5 mg or Placebo, DopAmine, Dorzolamide 2%/Timolol 0.5% Ophth., Dostinex, Doxazosin, Doxycycline Hyclate, Dronabinol, Duloxetine, Enalapril Maleate, Enalaprilat, Enoxaparin Sodium, Epidural Bag, Epoetin Alfa, Erythromycin, Erythromycin 0.5% Ophth Oint, Esmolol, Esmolol in Saline (Iso-osm), Etoposide, Exelon, Famotidine, Fat Emulsion 20%, Felodipine, Fentanyl Citrate, Fentanyl Patch, Ferrous Gluconate, Ferrous Sulfate, Ferrous Sulfate (Liquid), Filgrastim, Finasteride, Fleet Enema, Fleet Phospho-Soda, Fluconazole, Fludrocortisone Acetate, Fluocinonide 0.05% Ointment, Fluticasone Propionate 110mcg, Fluticasone Propionate NASAL, Fluticasone-Salmeterol (100/50), Fluticasone-Salmeterol (250/50) , Fluticasone-Salmeterol Diskus (100/50), Fluticasone-Salmeterol Diskus (250/50) , FoLIC Acid, Folic Acid, Fondaparinux Sodium, Fosinopril, Furosemide, Gabapentin, Gammagard Liquid, Gelclair, Gentamicin, GlipiZIDE, GlipiZIDE XL, Glipizide, Glucagon, Glycopyrrolate, Golytely, Guaifenesin, Guaifenesin-Dextromethorphan, HYDROmorphone (Dilaudid), HYDROmorphone P.F., Haloperidol, Heparin, Heparin Flush (10 units/ml), Heparin Flush (100 units/ml), Heparin Flush (5000 Units/mL), Heparin Flush CRRT (5000 Units/mL), Heparin Flush CVL (100 units/ml), Heparin Flush Hickman (100 units/ml), Heparin Flush PICC (100 units/ml), Heparin Lock Flush, Heparin Sodium, Hespan, Humulin-R Insulin, HydrALAZINE HCl, HydrALAzine, Hydralazine HCl, Hydrochlorothiazide, Hydrocodone-Acetaminophen, Hydrocortisone Na Succ., Hydrocortisone Na Succinate, Hydromorphone, Hyoscyamine, Ibuprofen Suspension, Imipramine, Indomethacin, Influenza Virus Vaccine, Insulin, Insulin Human Regular, Ipratropium Bromide MDI, Ipratropium Bromide Neb, Iso-Osmotic Dextrose, Iso-Osmotic Sodium Chloride, Isoniazid, Isosorbide Mononitrate (Extended Release), Isotonic Sodium Chloride, Ketorolac, LR, Labetalol, Labetalol HCl, Lactated Ringers, Lactulose, Lactulose Enema, Lansoprazole, Lansoprazole Oral Disintegrating Tab, Lansoprazole Oral Suspension, Leucovorin Calcium, Levalbuterol HCl, Levofloxacin, Levothyroxine Sodium, Lidocaine, Lidocaine 1%, Lidocaine 1%/Epinephrine 1:100000, Lidocaine 5% Patch, Lidocaine Jelly 2%, Lidocaine Jelly 2% (Urojet), Lidocaine Viscous 2%, Linezolid, Lisinopril, Loperamide, Lorazepam, Losartan Potassium, Lumigan, Maalox/Diphenhydramine/Lidocaine, Magnesium Citrate, Magnesium Oxide, Magnesium Sulfate, Mannitol 20%, Meclizine, Megestrol Acetate, Meperidine, Meropenem, Mesna, MetFORMIN XR (Glucophage XR), MetRONIDAZOLE (FLagyl), Metformin, Methadone, Methadone HCl, Methimazole, Methotrexate Sodium, Methotrexate Sodium P.F., MethylPREDNISolone Sodium Succ, Methylnaltrexone, Methylprednisolone, Methylprednisolone Na Succ, Metoclopramide, Metoprolol, Metoprolol Succinate XL, Metoprolol Tartrate, Metoprolol XL, Metronidazole, Miconazole Powder 2%, Midazolam, Midazolam HCl, Midodrine, Midodrine HCl, Milk of Magnesia, Milrinone, Mineral Oil, Minoxidil, Mirtazapine, Moexipril HCl, Montelukast Sodium, Morphine SR (MS Contin), Morphine Sulfate, Morphine Sulfate (Concentrated Oral Soln), Morphine Sulfate (Oral Soln.), Morphine Sulfate IR, Multivitamin 12, Multivitamin IV, Multivitamins, Multivitamins W/minerals, NIFEdipine CR, NORepinephrine, NS, NS (Glass Bottle), NS (Irrigation Bottle), NS (Mini Bag Plus), Nadolol, Nafcillin, Naloxone, Namenda, Naproxen, Neomycin-Polymyxin-Dexameth Ophth. Oint, Neostigmine, Nephrocaps, Neutra-Phos, NiCARdipine IV, Nicotine Patch, Nitroglycerin, Nitroglycerin Ointment 2%, Nitroglycerin SL, Nitroprusside Sodium, Norepinephrine, Nortriptyline, Nystatin, Nystatin Cream, Nystatin Ointment, Nystatin Oral Suspension, Nystatin-Triamcinolone Ointment, Octreotide Acetate, Olanzapine, Olanzapine (Disintegrating Tablet), Omeprazole, Ondansetron, Ondansetron ODT, Oxybutynin, OxycoDONE (Immediate Release) , OxycoDONE Liquid, OxycoDONE-Acetaminophen Elixir, Oxycodone, Oxycodone (Sustained Release), Oxycodone-Acetaminophen, Oxycodone-Acetaminophen Elixir, PHENYLEPHrine, PNEUMOcoccal Vac Polyvalent, Pancrelipase 5000, Pantoprazole, Pantoprazole Sodium, Papain 2.5 % Solution, Paroxetine, Paroxetine HCl, Penicillin G Potassium, Pentamidine-Inhalation, Phenazopyridine, Phenylephrine, Phenylephrine HCl, Phenytoin, Phenytoin Sodium, Phosphorus, Phytonadione, Pioglitazone HCl, Piperacillin-Tazobactam, Piperacillin-Tazobactam Na, Pletal, Pneumococcal Vac Polyvalent, Polyethylene Glycol, Potassium Chl 20 mEq / 1000 mL D5 1/2 NS, Potassium Chl 20 mEq / 1000 mL D5W, Potassium Chl 40 mEq / 1000 mL D5 1/2 NS, Potassium Chl 40 mEq / 1000 mL NS, Potassium Chloride, Potassium Chloride (Powder), Potassium Chloride Replacement (Oncology), Potassium Phosphate, Pravastatin, PredniSONE, PrednisoLONE Acetate 1% Ophth. Susp., Prednisone, Pregabalin, Prismasate (B22 K4), Prismasate (B32 K2), Prochlorperazine, Promethazine, Promethazine HCl, Propofol, Propofol (Generic), Propoxyphene Nap.-Apap (N-100), Propra, Propranolol, Pseudoephedrine HCl, Quetiapine Fumarate, Ranitidine, Rifaximin, Risperidone, Rituximab, Rosuvastatin Calcium, SW, Salmeterol Xinafoate Diskus (50 mcg), Sarna Lotion, Scopolamine Patch, Senna, Sertraline, Sevelamer, Simethicone, Simvastatin, Sirolimus, Sodium Bicarbo, Sodium Bicarbonate, Sodium CITRATE 4%, Sodium Chloride 0.9% Flush, Sodium Chloride 3% (Hypertonic), Sodium Chloride Nasal, Sodium Phosphate, Sodium Polystyrene Sulfonate, Soln, Soln., Solution, Spironolactone, Sterile Water, Sucralfate, Sulfameth/Trimethoprim, Sulfameth/Trimethoprim DS, Sulfameth/Trimethoprim SS, Sulfameth/Trimethoprim Suspension, Sumatriptan Succinate, Syringe, Syringe (0.9% Sodium Chloride), Syringe (Chemo), Tacrolimus, Tamsulosin, Thiamine, Thiamine HCl, Thioridazine HCl, Timolol Maleate 0.25%, Tiotropium Bromide, Tirofiban, Tizanidine, Tizanidine HCl, Tobramycin Sulfate, Tolterodine, Torsemide, TraMADOL (Ultram), Unasyn, Ursodiol, Valsartan, Vancomycin, Vancomycin Enema, Vancomycin HCl, Vancomycin Oral Liquid, Vasopressin, Venlafaxine XR, Verapamil SR, Vial, Vigamox, VinCRIStine (Oncovin), Viokase-8, Vitamin D, Warfarin, Xopenex, Xopenex Neb, Zinc Sulfate, Ziprasidone, Zolpidem Tartrate, sevelamer HYDROCHLORIDE, sodium bicar, traMADOL, traZODONE, traZODONE HCl, tucks}

1.4 Lab Tests

Here is the list of lab tests (MIMIC-III LABEVENTS/ITEMID) used in this study. { 50800, 50801, 50802, 50803, 50803, 50804, 50804, 50805, 50806, 50806, 50808, 50808, 50809, 50809, 50810, 50811, 50811, 50812, 50813, 50813, 50814, 50815, 50816, 50817, 50818, 50818, 50819, 50820, 50820, 50821, 50821, 50822, 50822, 50823, 50824, 50824, 50825, 50826, 50827, 50828, 50831, 50835, 50836, 50838, 50841, 50842, 50843, 50849, 50852, 50852, 50854, 50855, 50856, 50856, 50857, 50861, 50861, 50862, 50862, 50863, 50863, 50864, 50866, 50866, 50867, 50867, 50868, 50868, 50872, 50873, 50874, 50876, 50878, 50878, 50879, 50880, 50882, 50882, 50883, 50883, 50884, 50885, 50885, 50887, 50889, 50889, 50890, 50890, 50891, 50893, 50893, 50895, 50900, 50900, 50902, 50902, 50903, 50904, 50905, 50905, 50906, 50907, 50907, 50908, 50908, 50909, 50909, 50910, 50910, 50911, 50911, 50912, 50912, 50914, 50914, 50915, 50917, 50917, 50919, 50920, 50922, 50922, 50924, 50924, 50925, 50927, 50929, 50929, 50930, 50930, 50931, 50931, 50933, 50935, 50935, 50937, 50938, 50939, 50940, 50941, 50942, 50943, 50944, 50945, 50946, 50948, 50949, 50949, 50950, 50950, 50951, 50951, 50952, 50952, 50953, 50953, 50954, 50954, 50955, 50956, 50956, 50957, 50960, 50960, 50963, 50963, 50964, 50964, 50965, 50965, 50966, 50967, 50967, 50968, 50968, 50969, 50969, 50970, 50970, 50971, 50971, 50974, 50975, 50976, 50976, 50979, 50980, 50981, 50983, 50983, 50986, 50986, 50993, 50993, 50994, 50994, 50995, 50995, 50997, 50998, 50998, 50999, 51000, 51000, 51001, 51001, 51002, 51002, 51003, 51003, 51005, 51006, 51006, 51007, 51007, 51008, 51008, 51009, 51009, 51010, 51010, 51014, 51015, 51018, 51018, 51025, 51026, 51034, 51035, 51043, 51046, 51047, 51051, 51052, 51053, 51054, 51059, 51067, 51069, 51070, 51071, 51074, 51075, 51076, 51077, 51078, 51079, 51082, 51085, 51086, 51087, 51088, 51090, 51091, 51092, 51093, 51094, 51095, 51097, 51098, 51099, 51099, 51100, 51102, 51103, 51104, 51105, 51108, 51114, 51116, 51116, 51117, 51118, 51120, 51120, 51125, 51125, 51127, 51128, 51130, 51131, 51132, 51132, 51133, 51134, 51137, 51138, 51139, 51143, 51143, 51144, 51144, 51145, 51146, 51146, 51147, 51148, 51150, 51151, 51176, 51180, 51181, 51194, 51196, 51196, 51197, 51198, 51200, 51200, 51213, 51213, 51214, 51214, 51216, 51218, 51218, 51221, 51221, 51222, 51222, 51229, 51231, 51233, 51236, 51237, 51237, 51240, 51243, 51244, 51244, 51245, 51245, 51246, 51248, 51248, 51249, 51249, 51250, 51250, 51251, 51251, 51252, 51254, 51254, 51255, 51255, 51256, 51256, 51257, 51257, 51259, 51260, 51261, 51264, 51265, 51265, 51266, 51267, 51268, 51269, 51274, 51274, 51275, 51275, 51277, 51277, 51278, 51279, 51279, 51283, 51283, 51284, 51284, 51287, 51288, 51288, 51292, 51294, 51296, 51297, 51300, 51300, 51301, 51301, 51343, 51344, 51347, 51351, 51352, 51355, 51358, 51360, 51362, 51362, 51363, 51373, 51375, 51379, 51382, 51383, 51384, 51388, 51398, 51400, 51402, 51404, 51411, 51412, 51419, 51420, 51422, 51423, 51424, 51425, 51426, 51427, 51427, 51428, 51429, 51431, 51431, 51434, 51436, 51436, 51438, 51438, 51439, 51439, 51444, 51446, 51446, 51447, 51448, 51450, 51450, 51453, 51455, 51455, 51457, 51457, 51458, 51458, 51462, 51463, 51464, 51466, 51469, 51474, 51476, 51478, 51479, 51479, 51482, 51482, 51484, 51486, 51487, 51491, 51491, 51492, 51493, 51493, 51497, 51498, 51498, 51499, 51501, 51503, 51505, 51506, 51508, 51512, 51513, 51514, 51514, 51516, 51516, 51518, 51519, 51523'}

1.5 Microbiology Tests

Here is the list of microbiology test IDs used in this study: { 70002.0, 70003.0, 70005.0, 70006.0, 70008.0, 70009.0, 70010.0, 70011.0, 70012.0, 70013.0, 70014.0, 70018.0, 70020.0, 70021.0, 70022.0, 70023.0, 70024.0, 70026.0, 70027.0, 70028.0, 70029.0, 70030.0, 70031.0, 70033.0, 70034.0, 70036.0, 70037.0, 70038.0, 70040.0, 70041.0, 70042.0, 70047.0, 70048.0, 70050.0, 70051.0, 70053.0, 70054.0, 70057.0, 70058.0, 70059.0, 70061.0, 70062.0, 70063.0, 70064.0, 70067.0, 70068.0, 70069.0, 70070.0, 70072.0, 70074.0, 70075.0, 70076.0, 70079.0, 70080.0, 70082.0, 70083.0, 70084.0, 70085.0, 70086.0, 70090.0, 70091.0, 70092.0}

1.6 MeSH-based Symptoms

Here is the list of symptoms extracted from MIMIC-III NOTEEVENTS tables, utilizing MeSH representation. {abscess, abscesses, acalculia, acantholysis, ache, aches, agnosia, agnosias, agraphia, aivr, akathisia, albuminuria, alexia, allodynia, alopecia, amaurosis, amblyopia, amenorrhea, amnesia, amnesias, anasarca, aneuploidy, anhedonia, anisocoria, anomia, anorexia, anosmia, anosognosia, anoxia, aphasia, aphonia, apnea, apneas, apraxia, apraxias, arrhythmia, arrythmia, arthralgia, arthralgias, ascites, ascus, asphyxia, asterixis, asthenia, ataxia, atheroma, atheromas, athetosis, atrophies, atrophy, autolysis, azotemia, backache, backaches, bacteremia, bacteremias, baldness, belching, bilirubinemia, birthweight, bleb, blebs, bleeding, blindness, blister, blisters, bradyarrhythmia, bradyarrhythmias, bradycardia, bradycardias, bradykinesia, bradykinesias, breathlessness, bulimia, bulla, bullae, cachexia, cadasil, cadaver, calculi, calculus, candidemia, cardiomegaly, cardiotoxicities, cardiotoxicity, carsickness, catatonia, cellulites, cellulitis, cephalgia, cervicalgia, channelopathy, chills, chorea, cicatrix, cicatrization, cidp, cirrhosis, coad, coma, comas, comatose, confusion, constipation, convalescence, convulsion, convulsions, copd, cough, coughs, crackle, crackles, cramp, cramps, cyanosis, cyst, cystocele, cysts, deafness, death, debility, dehydration, delirium, diarrhea, diatheses, diathesis, diplopia, disease, diseases, disorientation, diverticula, diverticulosis, diverticulum, dizziness, dizzyness, drowning, dysarthria, dysarthrias, dyscalculia, dyschezia, dyscoordination, dysesthesia, dysesthesias, dysgeusia, dysgraphia, dyskinesia, dyskinesias, dyslexia, dysmenorrhea, dysmetria, dysnomia, dyspepsia, dysphasia, dysphonia, dyspnea, dyspraxia, dyssynergia, dystonia, dysuria, earache, earaches, ecchymoses, ecchymosis, echolalia, ectasia, edema, emergencies, emergency, emesis, emphysema, empyema, encephalocele, encopresis, endoleak, endoleaks, enterocele, epistaxis, eructation, erythema, erythemas, esrd, exsanguination, extrasystole, extrasystoles, facies, fainting, fasciculation, fasciculations, fatigue, fever, fevers, fibroses, fibrosis, fistula, fistulas, flatulence, flatus, flushing, flushings, formication, formications, frailty, fungemia, gagging, gallstone, gallstones, gangrene, gastroparesis, gastroschisis, gegenhalten, gliosis, granuloma, granulomas, haemolysis, halitosis, hallucination, hallucinations, headache, headaches, heartburn, hemarthroses, hemarthrosis, hematemesis, hematocele, hematoceles, hematochezia, hematoma, hematomas, hematuria, hemianopia, hemianopsia, hemiballismus, hemicrania, hemiparesis, hemiplegia, hemobilia, hemoglobinuria, hemolysis, hemoperitoneum, hemopneumothorax, hemoptyses, hemoptysis, hemorrhage, hemorrhages, hemothorax, hepatomegaly, hernia, hernias, hiccoughs, hiccup, hiccups, hirsutism, hoarseness, hydrops, hyperacusis, hyperalgesia, hyperammonemia, hyperamylasemia, hyperbilirubinemia, hypercapnia, hyperesthesia, hyperesthesias, hypergammaglobulinemia, hyperkinesis, hyperlactatemia, hypermetria, hyperoxia, hyperplasia, hyperpyrexia, hyperreflexia, hyperthermia, hypertrophy, hyperuricemia, hyperventilation, hypesthesia, hyphema, hypoacusis, hypocapnia, hypokinesia, hypomania, hypomenorrhea, hyporeflexia, hypothermia, hypotonia, hypoventilation, hypoventilations, hypovolemia, hypovolemic, hypoxemia, hypoxia, icterus, illusion, illusions, incoordination, indigestion, infarct, infarction, infarctions, infarcts, inflammation, ischemia, isochromosome, itching, jaundice, keloid, keloids, lethargy, leukoaraiosis, leukocytosis, leukoplakia, lightheadedness, lithiasis, lumbago, macroamylasemia, mania, melena, melenas, meningismus, meningocele, meningoceles, menorrhagia, metaplasia, metastases, metastasis, metrorrhagia, micrometastases, miosis, mods, monosomy, mutism, myalgia, myoclonus, myokymia, myotoxicity, narcosis, nausea, neckache, necrosis, neointima, nephrolith, neuralgia, neuralgias, nocturia, nosebleed, nosebleeds, nstemi, numbness, obesity, oligomenorrhea, oliguria, omphalocele, ophthalmoplegia, orthodeoxia, orthopnea, orthostasis, osteonecrosis, otalgia, ototoxicity, overeating, overweight, pain, pallor, palsies, palsy, paralyses, paralysis, paraparesis, paraplegia, parasitemia, paresis, paresthesia, paresthesias, petechiae, phlegmon, phonophobia, photophobia, pica, platypnea, plegia, pleocytosis, pneumomediastinum, pneumoretroperitoneum, polyarthralgia, polyarthralgias, polydipsia, polymyoclonus, polyp, polyphagia, polyploid, polyps, polyuria, presbycusis, presyncope, prolapse, prolapses, prosopagnosia, prostatism, proteinuria, pruritis, pruritus, pseudophakia, purpura, pus, pyemia, pyothorax, pyrexia, quadrantanopia, quadrantanopsia, quadriparesis, quadriplegia, rale, rales, recrudescence, rectocele, recurrence, recurrences, reinfection, relapse, relapses, restlessness, reticulocytosis, rhinorrhea, rhonchi, rhonchus, scar, scarring, scars, sciatica, sclerosis, scotoma, scotomas, seizure, seizures, sepsis, septicemia, septicemias, seroma, seromas, serositis, shock, sialolith, sialolithiasis, sialoliths, sid, sids, sleepiness, sneezing, snoring, somnolence, spasm, spasms, spastic, spermatocele, splenomegaly, spotting, stammering, stds, stemi, stenoses, stenosis, sti, stillbirth, stis, stricture, strictures, stridor, stridors, stunting, stupor, stuttering, suffocation, suppuration, syncope, syncopes, syndrome, syndromes, synkinesis, tachyarrhythmia, tachyarrhythmias, tachycardia, tachycardias, tachypnea, tetany, tetraplegia, tetraploid, tetrasomy, thinness, tic, tics, tinnitus, toothache, toothaches, torticollis, tremor, tremors, trismus, trisomy, ulcer, ulcers, unconsciousness, underweight, urinoma, vasoplegia, vertigo, viremia, virilization, volvulus, vomiting, wheezing}.

2 Methodology and Implementation

2.1 Learning Models

2.1.1 ANN

Artificial Neural Networks (ANNs)[1,2] are machine learning models structured as layers of interconnected neurons, with each layer transforming the input data into higher-level representations. A typical form of ANN is the *Multi-Layer Perceptron (MLP)*, composed of an input layer, one or more hidden layers, and an output layer. Mathematically, the output of each layer is computed as:

$$\mathbf{h}^{(l)} = f(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}),$$

where $\mathbf{h}^{(l)}$ is the activation of layer l, $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ are the weight matrix and bias vector of layer l, and $f(\cdot)$ is the activation function (e.g., ReLU, $f(x) = \max(0,x)$) applied element-wise.

In a supervised learning setup, the MLP minimizes a loss function L over the output predictions $\hat{\mathbf{y}}$ compared to the ground truth labels $\mathbf{y}$. For multi-class classification, the softmax function is applied at the output layer to compute class probabilities:

$$\hat{y}_i = \frac{\exp(z_i)}{\sum_{j=1}^{C}\exp(z_j)},$$

where z_i is the output logit for class i, and C is the number of classes. The loss is typically the cross-entropy:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}),$$

where N is the number of samples.

Training an MLP involves optimizing weights and biases through gradient descent techniques, with gradients computed via backpropagation. These gradients are propagated layer by layer, leveraging the chain rule of differentiation, enabling efficient updates of parameters. MLPs excel at capturing non-linear relationships in structured data and are widely used for tasks such as classification, regression, and representation learning in various domains.

2.1.2 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost)[3] is an optimized implementation of gradient boosting that is widely used for classification and regression tasks. XGBoost builds an ensemble of decision trees, where each tree attempts to correct the errors of its predecessor. For a multi-class classification task, XGBoost minimizes a regularized objective function, defined as:

$$\mathscr{L}(\Theta) = \sum_{i=1}^{N} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right),$$

where $l(y_i, \hat{y}_i)$ is the loss function (e.g., multi-class log loss), $\Omega(f_k)$ is the regularization term for tree complexity, N is the number of samples, and K is the number of trees. Each tree f_k contributes to the prediction $\hat{y}_i$ as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i),$$

where f_k is a decision tree, and $\mathbf{x}_i$ represents the feature vector for sample i.

During training, XGBoost employs a second-order Taylor expansion of the loss function to approximate gradients and optimize the objective efficiently. The split criterion for a tree is determined by maximizing the gain G, which is given as:

$$G = \frac{1}{2}\left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda}\right],$$

where g_i and h_i are the first and second-order gradients for sample i, I_L and I_R are the left and right child nodes, and λ is a regularization parameter.

2.1.3 Random Forest

Random Forest[4] is an ensemble learning method widely used for classification and regression tasks. It operates by constructing multiple decision trees during training and aggregating their predictions for improved accuracy and robustness. For a given dataset $(\mathbf{X}, \mathbf{Y})$, Random Forest builds N decision trees by repeatedly sampling the data with replacement (bootstrap sampling) and selecting a subset of features for each tree. The prediction for classification is determined through majority voting across the individual trees.

The construction of each decision tree involves splitting the input space to maximize information gain or minimize impurity. The impurity of a split is typically measured using metrics like Gini impurity[5]:

$$G = \sum_{i=1}^{C} p_i(1 - p_i),$$

where p_i is the proportion of samples belonging to class i in the split, and C is the number of classes. The tree-building process continues until a stopping criterion, such as maximum depth or minimum samples per leaf, is reached.

For multi-class classification, the final prediction for an input $\mathbf{x}$ is:

$$\hat{y} = \text{argmax}_c \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(T_i(\mathbf{x}) = c),$$

where $T_i(\mathbf{x})$ is the class predicted by the i-th tree, and $\mathbb{I}$ is an indicator function.

2.1.4 Logistic Regression

Logistic Regression[6] is a linear model commonly used for binary and multi-class classification tasks. It models the relationship between input features $\mathbf{X} \in \mathbb{R}^{N \times d}$ and the target variable $\mathbf{y} \in \{1, \ldots, C\}^N$ by estimating the probability of each class using the softmax function. For a given input $\mathbf{x}_i$, the predicted probabilities for class j are:

$$P(y = j \mid \mathbf{x}_i) = \frac{\exp(\mathbf{w}_j^\top \mathbf{x}_i + b_j)}{\sum_{k=1}^{C} \exp(\mathbf{w}_k^\top \mathbf{x}_i + b_k)},$$

where $\mathbf{w}_j \in \mathbb{R}^d$ is the weight vector, b_j is the bias term, and C is the number of classes. Logistic Regression optimizes the cross-entropy loss function to maximize the likelihood of the predicted probabilities:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(P(y=j \mid \mathbf{x}_i)),$$

where $y_{ij} \in \{0,1\}$ is a one-hot encoded indicator for the true class of sample i.

For multi-class classification, the one-vs-rest (OvR) strategy is used, where the model independently learns a binary classifier for each class against all others. During training, weights $\mathbf{w}$ and biases $\mathbf{b}$ are optimized iteratively using solvers such as `lbfgs`, which efficiently handle high-dimensional data and ensure convergence.

2.1.5 Support Vector Machines

Support Vector Machines (SVMs)[7] are powerful supervised learning models used for classification and regression tasks. For multi-class classification, the one-vs-rest (OvR) strategy is commonly employed, where an SVM is trained to separate each class from the others. The SVM finds the optimal hyperplane that maximizes the margin between classes in the feature space. The decision boundary is defined by:

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b,$$

where $\mathbf{w}$ is the weight vector, b is the bias, and $\mathbf{x}$ is the input feature vector. For non-linearly separable data, SVMs utilize kernel functions $K(\mathbf{x}_i, \mathbf{x}_j)$ to project data into a higher-dimensional space, enabling linear separability. The commonly used Radial Basis Function (RBF) kernel is defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right),$$

where γ controls the kernel's sensitivity to data points.

The SVM optimizes the following objective function to minimize classification error and maximize the margin:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i,$$

subject to the constraints $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i$, where $\xi_i \geq 0$ are slack variables allowing some misclassification, and C is a regularization parameter balancing margin maximization and classification error.

2.2 Implementation details

We incorporate the following settings for the machine learning models. A Random Forest with $N = 250$ estimators was employed to capture non-linear relationships through robust ensemble learning. Logistic Regression, using the `lbfgs` solver, was applied with standardized features for scalable classification. Support Vector Machines (SVMs) were configured with a radial basis function (RBF) kernel and enabled probabilistic outputs for multi-class tasks. An LSTM model, designed with 128 hidden units, was trained for sequential data over 50 epochs using the Adam optimizer. Finally, XGBoost was optimized with a `multi:softprob` objective, a learning rate of 0.1, a maximum depth of 6, and 250 boosting rounds, incorporating early stopping to prevent overfitting.

2.3 Data Cleaning

Structured Data Representation To create one-hot encoding for each hospital visit, we generate binary feature vectors that represent the presence or absence of clinical items across specific categories: Diagnoses (D), Procedures (P), Medications (M), Laboratory tests (L), and Microbiology results (B). For each visit, a binary value of 1 is assigned if a clinical item within a given category exists for that visit, and 0 otherwise. This encoding ensures that each category is represented as a separate vector, capturing whether any clinical item was documented during the visit. By consolidating these vectors, we create a comprehensive representation of the clinical events that occurred during a visit, facilitating analysis and model input in machine learning tasks.

Unstructured Data Representation The process begins by segmenting the unstructured clinical notes into key sections, such as chief complaints, medications, physical examinations, and family history, to capture structured and relevant clinical information for each hospital visit. These sections are then concatenated to form a unified textual representation per visit, encompassing the breadth of clinical details.

2.4 Classification Metrics

The evaluation of model performance was conducted using standard classification metrics: *accuracy*, *precision*, *recall*, *F1-score*, and *area under the ROC curve (AUC)*.

- **Accuracy** measures the proportion of correctly classified instances among all predictions:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

- **Precision** (Positive Predictive Value) quantifies the proportion of true positives among all positive predictions:

$$\text{Precision} = \frac{TP}{TP+FP}$$

- **Recall** (Sensitivity or True Positive Rate) represents the proportion of actual positives that were correctly identified:

$$\text{Recall} = \frac{TP}{TP+FN}$$

- **F1-score** is the harmonic mean of precision and recall, balancing both false positives and false negatives:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **AUC** (Area Under the Receiver Operating Characteristic Curve) summarizes the trade-off between true positive rate and false positive rate across thresholds. A higher AUC indicates stronger discriminative capability.

Here, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. In multi-class classification, precision, recall, and F1-score are computed for each class in a one-vs-rest scheme and then aggregated using macro, micro, or weighted averages. The area under the curve (AUC) can be generalized using one-vs-rest or pairwise strategies.

3 Results

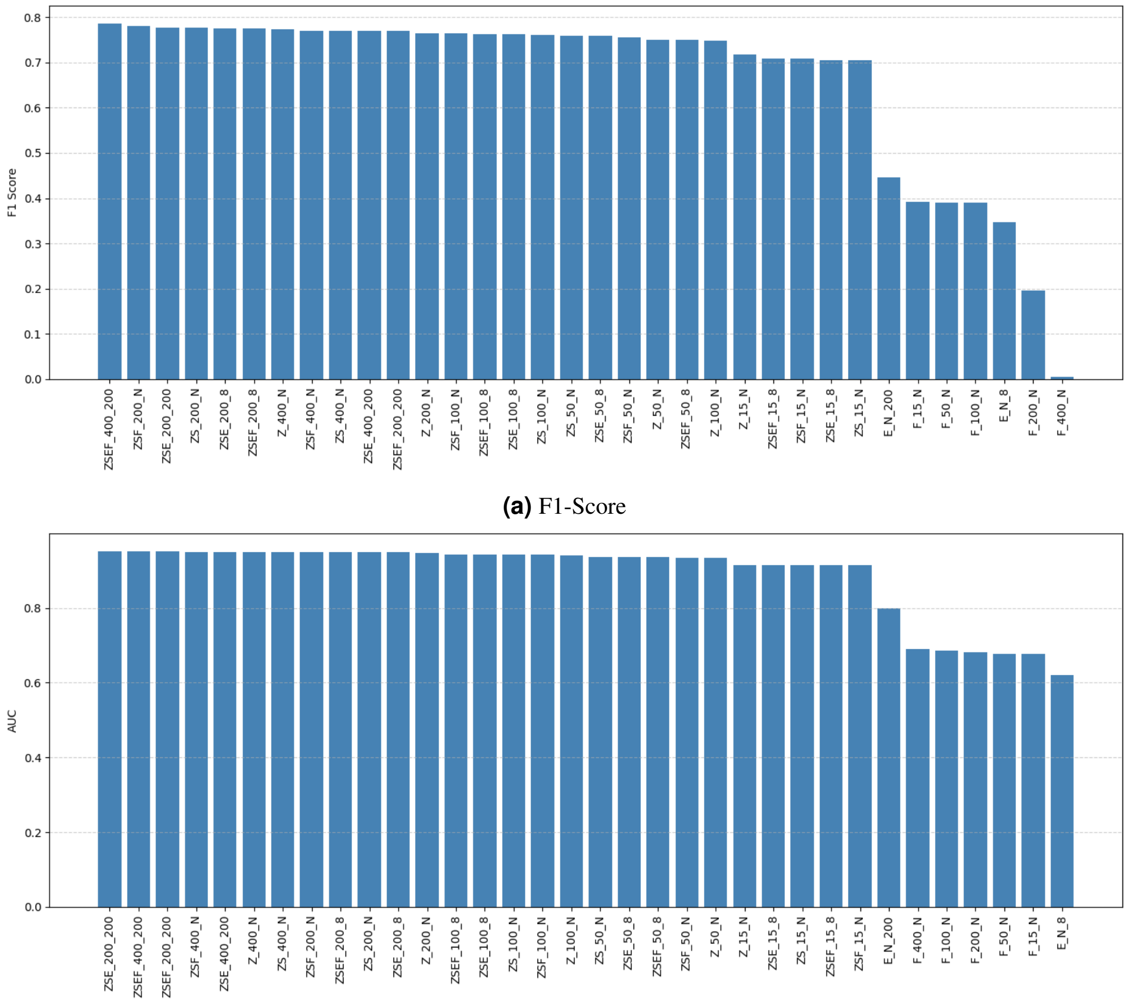
3.1 Performance Evaluation Using Feature Selection on 2 Classes

An intensive analysis is conducted to evaluate the impact of the feature selection on the performance of the ANN model. Herein, we used (15, 50, 100, 200, and 400) top features from the categorical component of the data ($Z = [D, P, M, L, B], S, F$), and (0, 8, 200) as **N**one, low, and high top features from the BoW representation (E). Table 1 shows different evaluation metrics used for this use case. Figure 1 shows the performance of the different settings in descending order (F1-score (Fig. 1a), AUC (Fig. 1b)).

Performance Analysis. Feature selection had a measurable effect on ANN model performance, as detailed in Figure 1. The best-performing configuration (`ZSEF_400_200`, F1 = 0.7864) approached the baseline performance without selection (F1 = 0.79), confirming the robustness of well-selected features. Notably, the Z feature set consistently demonstrated strong predictive capacity across configurations, with incremental gains from the inclusion of symptoms data S. In contrast, contributions from E (BoW) and F (physiological features) were marginal and statistically insignificant. Configurations relying primarily on F or E showed poor performance (e.g., `F_200_N`, F1 = 0.0024), underscoring their limited standalone utility. The average performance of ZSF or ZSE combinations suggests that moderate feature reduction can maintain model accuracy, provided core structured features are preserved.

3.2 Class Sensitivity Analysis

To further understand the effect of class granularity on model performance, we conducted a sensitivity analysis by varying the number and boundaries of length of stay (LoS) classes. This analysis helps reveal how classification difficulty and feature effectiveness change as the prediction task becomes more fine-grained.



(a) F1-Score

(b) AUC

Figure 1. Impact of the different feature selection sets on the performance of the different data types (Z, F, E, ZS, ZSF, ZSE, ZSEF) using ANN. Each bar reflects a combination of (data used, number of features used for the categorical component, and number of features used for BoW). **N** means no BoW embedding is presented in the given data.

3.2.1 Three Classes (3, 5)

In this setting, hospital stays were categorized into three classes: short ($1 \leq \text{LoS} \leq 3$ days), medium ($3 < \text{LoS} \leq 5$ days), and long ($\text{LoS} > 5$ days). This configuration introduces a narrower medium stay range. Performance metrics for this setup are presented in Table 2.

3.2.2 Four Classes (3, 7, 30)

Here, we further subdivided long stays into four categories to reflect greater clinical variation. The four classes are defined as: short ($1 \leq \text{LoS} \leq 3$ days), medium ($3 < \text{LoS} \leq 7$ days), long ($7 < \text{LoS} \leq 30$ days), and prolonged ($\text{LoS} > 30$ days). This more granular stratification enables evaluation of model sensitivity to rare but clinically important prolonged stays. Detailed performance outcomes are summarized in Table 3.

Acknowledgment

References

1. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin mathematical biophysics* **5**, 115–133 (1943).

2. Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. review* **65**, 386 (1958).

3. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).

4. Liaw, A., Wiener, M. *et al.* Classification and regression by randomforest. *R news* **2**, 18–22 (2002).

5. Pal, M. Random forest classifier for remote sensing classification. *Int. journal remote sensing* **26**, 217–222 (2005).

6. Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: a methodology review. *J. biomedical informatics* **35**, 352–359 (2002).

7. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. learning* **20**, 273–297 (1995).

Table 1. Performance of ANN across various data configurations and feature selection dimensions. FS_structured and FS_E represent the implementation of feature selection models for each of the structured and unstructured data types, respectively. The N value in FS_E column means no note embedding (E) is incorporated in the corresponding data type.

DataType	FS_Structured	FS_E	Accuracy	Precision	Recall	F1 Score	AUC
ZSEF	400	200	0.9155	0.8023	0.7712	**0.7864**	0.9521
ZSEF	200	200	0.9156	0.8586	0.6965	0.7691	0.9513
ZSEF	200	8	0.9147	0.8268	0.7300	0.7754	0.9492
ZSEF	100	8	0.9105	0.8213	0.7112	0.7623	0.9431
ZSEF	50	8	0.9065	0.8158	0.6929	0.7494	0.9356
ZSEF	15	8	0.8955	0.8095	0.6300	0.7086	0.9140
ZSF	400	N	0.9121	0.8132	0.7324	0.7707	0.9503
ZSF	200	N	0.9126	0.7808	0.7876	0.7842	0.9498
ZSF	200	N	0.9152	0.8312	0.7271	0.7757	0.9489
ZSF	100	N	0.9108	0.8168	0.7188	0.7647	0.9423
ZSF	50	N	0.9062	0.7973	0.7171	0.7550	0.9350
ZSF	15	N	0.8937	0.7934	0.6394	0.7081	0.9139
ZSE	400	200	0.9129	0.8233	0.7235	0.7702	0.9497
ZSE	200	200	0.9162	0.8390	0.7235	0.7770	**0.9522**
ZSE	200	8	0.9147	0.8246	0.7329	0.7761	0.9488
ZSE	100	8	0.9101	0.8174	0.7135	0.7619	0.9430
ZSE	50	8	0.9081	0.8049	0.7182	0.7591	0.9359
ZSE	15	8	0.8951	0.8153	0.6206	0.7047	0.9142
ZS	400	N	0.9134	0.8273	0.7212	0.7706	0.9495
ZS	200	N	0.9145	0.8104	0.7518	0.7800	0.9497
ZS	200	N	0.9128	0.8123	0.7382	0.7735	0.9481
ZS	100	N	0.9101	0.8200	0.7100	0.7610	0.9428
ZS	50	N	0.9035	0.7648	0.7535	0.7591	0.9363
ZS	15	N	0.8948	0.8120	0.6224	0.7046	0.9141
Z	400	N	0.9138	0.8207	0.7324	0.7740	0.9496
Z	200	N	0.9120	0.7818	0.7818	0.7818	0.9489
Z	200	N	0.9102	0.8602	0.6624	0.7484	0.9479
Z	100	N	0.9089	0.8447	0.6718	0.7484	0.9419
Z	50	N	0.9035	0.7841	0.7200	0.7507	0.9337
Z	15	N	0.8962	0.7953	0.6535	0.7175	0.9143
F	400	N	0.7981	0.4000	0.0024	0.0047	0.6904
F	200	N	0.8214	0.6265	0.2841	0.3909	0.6796
F	200	N	0.7986	0.6231	0.0012	0.0024	0.6842
F	100	N	0.8206	0.6211	0.2835	0.3893	0.6852
F	50	N	0.8214	0.6265	0.2841	0.3909	0.6774
F	15	N	0.8181	0.6020	0.2900	0.3914	0.6772
E	N	200	0.8341	0.6824	0.3324	0.4470	0.7994
E	N	200	0.8351	0.6937	0.3265	0.4440	0.7987
E	N	8	0.8295	0.7604	0.2259	0.3483	0.6211
E	N	8	0.8293	0.7584	0.2253	0.3474	0.6210
E	N	8	0.8293	0.7594	0.2247	0.3468	0.6211
E	N	8	0.8293	0.7615	0.2235	0.3456	0.6211

All Precision, Recall, F1-score, and AUC values are computed as weighted averages to account for class imbalance.

Table 2. Three classes performance metrics of different Machine Learning models and data types. The three classes are as follow: short ($1 \leq \text{LoS} \leq 3$ days), medium ($3 < \text{LoS} \leq 5$ days), and long ($\text{LoS} > 5$ days). No feature selection (FS) models are implemented.

ML Model	Data Type	Accuracy	Precision	Recall	F1 Score	AUC
ANN	ZSF	0.7451	0.6901	0.7451	0.6903	**0.8631**
ANN	ZSEF	0.7484	0.6790	0.7484	0.6862	0.8630
ANN	ZSE	0.6948	0.6946	0.6948	0.6946	0.8375
ANN	ZS	0.6929	0.7015	0.6929	0.6965	0.8339
ANN	Z	0.7166	0.7006	0.7166	**0.7057**	0.8403
ANN	F	0.5705	0.5184	0.5705	0.5265	0.6309
ANN	E	0.6312	0.5856	0.6312	0.6011	0.7342
XGBoost	ZSF	**0.7491**	0.6967	**0.7491**	0.6947	0.8436
XGBoost	ZSEF	0.7480	0.6923	0.7480	0.6923	0.8457
XGBoost	ZSE	0.7482	0.6957	0.7482	0.6936	0.8446
XGBoost	ZS	0.7474	0.6913	0.7474	0.6934	0.8454
XGBoost	Z	0.7488	0.7004	0.7488	0.6946	0.8428
XGBoost	F	0.6140	0.5585	0.6140	0.5314	0.6465
XGBoost	E	0.6673	0.6283	0.6673	0.5985	0.7273
LR	ZSF	0.7334	0.6749	0.7334	0.6851	0.8318
LR	ZSEF	0.7365	0.6825	0.7365	0.6880	0.8321
LR	ZSE	0.7354	0.6932	0.7354	0.7051	0.8288
LR	ZS	0.7379	0.6902	0.7379	0.7011	0.8301
LR	Z	0.7393	0.6897	0.7393	0.6992	0.8326
LR	F	0.6043	0.4973	0.6043	0.5192	0.6145
LR	E	0.6640	0.6093	0.6640	0.6040	0.7242
RF	ZSF	0.7379	0.6845	0.7379	0.6669	0.8275
RF	ZSEF	0.7325	**0.71**	0.7325	0.6618	0.8150
RF	ZSE	0.7288	0.6799	0.7288	0.6567	0.8111
RF	ZS	0.7372	0.6801	0.7372	0.6669	0.8234
RF	Z	0.7368	0.6806	0.7368	0.6688	0.8263
RF	F	0.6149	0.5589	0.6149	0.5297	0.6451
RF	E	0.6514	0.5849	0.6514	0.5747	0.7077
SVM	ZSF	0.2669	0.502	0.2669	0.1173	0.7195
SVM	ZSEF	0.2818	0.482	0.2818	0.1514	0.6473
SVM	ZSE	0.5097	0.5453	0.5097	0.5228	0.7061
SVM	ZS	0.5175	0.5446	0.5175	0.5266	0.7007
SVM	Z	0.5293	0.5679	0.5293	0.535	0.7161
SVM	F	0.2765	0.4602	0.2765	0.1577	0.5169
SVM	E	0.325	0.4901	0.322	0.2067	0.5703

All Precision, Recall, F1-score, and AUC values are computed as weighted averages to account for class imbalance.

Table 3. Four classes performance metrics of different Machine Learning models and data types. The four classes are as follow: short ($1 \leq \text{LoS} \leq 3\,\text{days}$), medium ($3 < \text{LoS} \leq 7\,\text{days}$), long ($7 < \text{LoS} \leq 30\,\text{days}$), and prolong long ($\text{LoS} > 30\,\text{days}$). No feature selection (FS) models are implemented.

ML Model	Data Type	Accuracy	Precision	Recall	F1 Score	AUC
ANN	ZSEF	0.7174	0.6899	0.7174	0.6873	0.8547
ANN	ZSF	0.7138	0.6857	0.7138	0.6773	0.8510
ANN	ZSE	0.6810	0.6890	0.6810	0.6830	0.8334
ANN	ZS	0.6753	0.6786	0.6753	0.6766	0.8295
ANN	Z	0.7023	0.6863	0.7023	0.6889	0.8351
ANN	F	0.5592	0.4820	0.5592	0.4181	0.6299
ANN	E	0.5885	0.5598	0.5885	0.5694	0.7187
LR	ZSEF	0.7053	0.6801	0.7053	0.6841	0.8753
LR	ZSF	0.7049	0.6815	0.7049	0.6861	0.8728
LR	ZSE	0.7117	0.6916	0.7117	**0.6974**	0.8791
LR	ZS	0.7104	0.6878	0.7104	0.6936	0.8776
LR	Z	0.7122	0.6889	0.7122	0.6941	0.8810
LR	F	0.5750	0.4698	0.5750	0.4644	0.6567
LR	E	0.6166	0.5713	0.6166	0.5693	0.7671
RF	ZSEF	0.6887	0.6481	0.6887	0.6288	0.8634
RF	ZSF	0.7050	0.6708	0.7050	0.6609	0.8761
RF	ZSE	0.6861	0.6415	0.6861	0.6270	0.8628
RF	ZS	0.7071	0.6762	0.7071	0.6649	0.8758
RF	Z	0.7042	0.6724	0.7042	0.6635	0.8765
RF	F	0.5814	0.4877	0.5814	0.4759	0.6792
RF	E	0.6048	0.5389	0.6048	0.5139	0.7443
XGBoost	ZSEF	0.7204	**0.6943**	0.7204	0.6943	0.8903
XGBoost	ZSF	0.7189	0.6923	0.7189	0.6944	**0.8907**
XGBoost	ZSE	**0.7205**	0.6939	**0.7205**	0.6943	0.8895
XGBoost	ZS	0.7182	0.6915	0.7182	0.6934	0.8901
XGBoost	Z	0.7160	0.6886	0.7160	0.6910	0.8885
XGBoost	F	0.5809	0.4914	0.5809	0.4820	0.6878
XGBoost	E	0.6241	0.5717	0.6241	0.5623	0.7703

All Precision, Recall, F1-score, and AUC values are computed as weighted averages to account for class imbalance.